UNIVERSIDAD CARLOS III DE MADRID

SISTEMA AUTOMÁTICO DE GENERACIÓN DE TENDENCIAS ARTIFICIALES.

Autor: Alberto Chicharro Sobrino

Tutor: Juan Miguel Carrascosa Amigo

Co-Tutor: Rubén Cuevas Rumin

Contenido

Índice de Ilustraciones:

1 Introducción

Con el aumento de popularidad que han sufrido las redes sociales en los últimos años, estas se han convertido en una herramienta indispensable, tanto como para el uso particular como para el empresarial. Entre los casos más representativos tenemos a la red social Twitter, que con su carácter de blog, donde los mensajes están limitados a 140 caracteres, se ha convertido en una de las más utilizadas junto a Facebook.

En la actualidad casi todo el mundo tiene una cuenta en alguna red social, cuando no en varias, por lo que las redes sociales se han vuelto en una herramienta indispensable en nuestras vidas para mantener el contacto con otras personas.

Teniendo esto en cuenta, las redes sociales ofrecen una oportunidad inmejorable para las empresas de proporcionar publicidad de una forma sencilla y cercana al usuario.

En este proyecto se desarrollará una herramienta capaz de utilizar cuentas de Twitter para publicar contenido en la red utilizando un mismo mensaje de tal forma que este se encuentre entre las tendencias más populares del momento. Dentro de esta herramienta se podrá elegir como cada una de las cuentas involucradas dentro del experimento envía mensajes a la red, permitiendo la personalización de la distribución que se desea generar. Además, de elegir la distribución con la que se publicarán los mensaje se podrá elegir la ubicación en la cual estos estarán vinculados, de tal forma que tendremos una mayor libertad de uso.

Para realizar este proyecto nos estamos apoyando en dos de las herramientas que utiliza Twitter, el etiquetado de palabras, también conocido con el nombre de hashtag, hace que las frases que contienen esta palabra sean más sencillas de encontrar y los Trending Topic, que muestran las tendencias más populares en una zona geográfica determinada.

1.1.Objetivos

Una vez hemos definido brevemente en qué consiste nuestra herramienta vamos a hablar sobre los objetivos que nos gustaría cumplir con ella. Estos son dos esencialmente.

El principal objetivo del proyecto es la generación de una tendencia artificial. Para conseguir este objetivo deberemos tener cuidado con nuestras cuentas vinculadas a la herramienta, ya que al publicar mucho contenido durante un periodo de tiempo relativamente corto, pueden llegar a ser etiquetadas como spam. Para evitar este posible escenario utilizaremos un sistema generador de lenguaje, que nos permitirá modificar el contenido que publica cada uno de los usuarios. Además, para facilitar lograr el objetivo deseado, se utilizará la geolocalización proporcionada por Twitter, de tal forma que podamos lograr ser una de las tendencias más influyentes en un área concreta de la geografía española.

El segundo objetivo del proyecto será el análisis de población activa en la red social de Twitter en cada una de las zonas geográficas en las que Twitter divide España. Para ello, se realizará un examen exhaustivo de las cuentas necesarias para conseguir aparecer en cada una de las zonas geográficas. Con este segundo objetivo se pretende identificar como la población interactúa dentro de las redes sociales y el grado de actividad de los usuarios en la red, en particular en Twitter.

Además, realizando los experimentos correspondientes a cada uno de los objetivos podremos identificar el número de personas que serían necesarias para generar una tendencia en cualquier geolocalización, no solo en las propuestas por nuestra herramienta.

1.2.Motivaciones

Como se ha dicho previamente, el auge de las redes sociales desde su creación ha ido en constante crecimiento, siendo empresas que incluso cotizan en Bolsa [36,37]. El gran aumento de usuarios que estas han experimentado desde su creación ha hecho que las redes sociales se conviertan en uno de los principales focos de la sociedad, siendo una de las formas más eficientes y cómodas de comunicación.

Las herramientas que Twitter ofrece tanto a usuarios como desarrolladores, ha permitido que Twitter sea una de las redes sociales más utilizadas en la actualidad.

Como se comento en la sección anterior, uno de estos objetivos era ser capaz mediante nuestra herramienta de generar un Trending Topic artificial, de tal forma que el hashtag utilizado en el experimento será visto por todos los usuarios activos del momento en la zona geográfica en la que se ha estado publicando. De esta forma, con esta herramienta conseguiríamos que decenas o cientos de miles de personas, dependiendo de la geolocalización seleccionada, vean nuestro hashtag como uno de los temas más hablados.

Para lograr este objetivo debemos conocer cómo funcionan los Trending Topic en Twitter, para que la herramienta sea capaz de generar el suyo propio con el menor esfuerzo posible y de la manera más eficiente. Para ello, deberemos analizar el comportamiento de los usuarios dentro de las redes sociales así como el propio algoritmo que Twitter utiliza para determinar que mensajes son los más influyentes en un determinado momento.

Además de lo previamente comentado, el gran volumen de gente involucrada dentro de las redes sociales ofrece una oportunidad inigualable a las empresas para ofrecer su publicidad.

En los últimos años, muchas de las redes sociales han ido añadiendo publicidad en sus páginas, ya que al tratarse de un servicio gratuito, las redes sociales no tenían ningún tipo de ingreso. Con la agregación de esta publicidad, las redes sociales mediante la información que los usuarios proporcionan en sus perfiles, proporcionan una publicidad personalizada como la que ofrecen amazon o gmail en sus propias páginas.

Entre las redes sociales más conocidas que han implementado un sistema de publicidad propia se encuentran Facebook y Twitter, que son las dos con un mayor número de usuarios activos.

De esta forma, si nuestra herramienta es capaz de generar un Trending Topic mediante el uso de cuentas virtuales, el hashtag utilizado para ser Trending Topic seria visible para todos los usuarios activos en la zona geográfica en la que este resulte tendencia. Por tanto, esta herramienta ofrecería una oportunidad de generar publicidad dentro de Twitter sin ningún tipo coste.

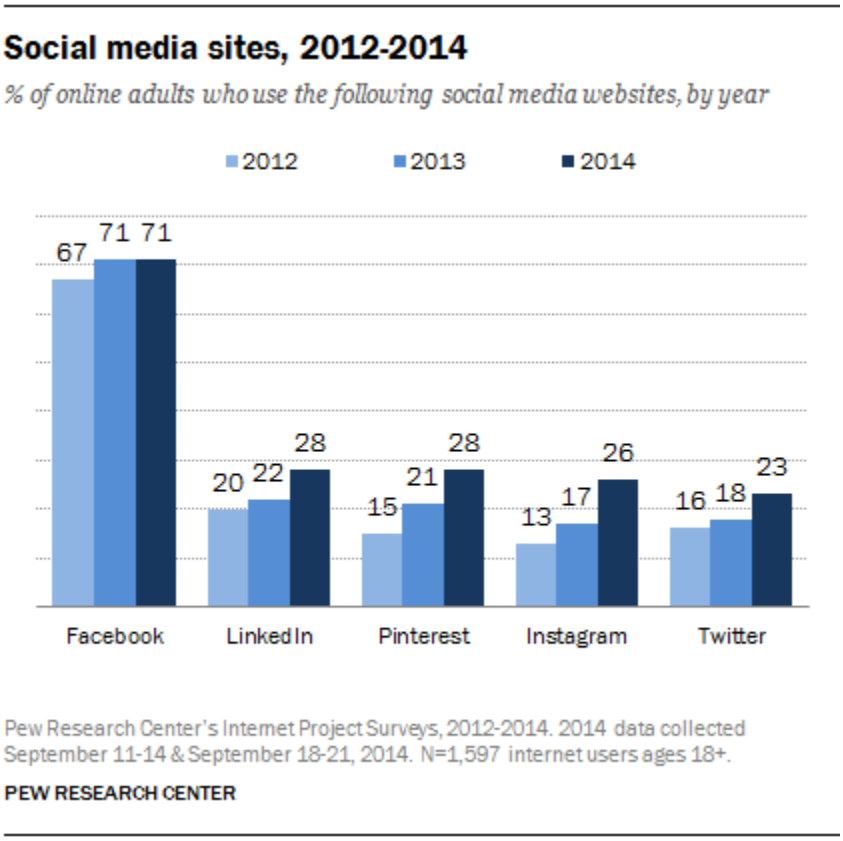
2. Redes sociales

2.1.¿Qué es una red social?

Una red social es una estructura social formada por personas conectadas y unidas entre sí por algún tipo de relación y/o interés común [1,2]. Estas son parte de nuestra vida, ya que es la forma en que nos relacionamos con las personas. En el estudio de redes sociales estas se suelen representar mediante grafos, conectando las personas que tienen alguna relación entre ellas.

En los últimos años, con la llegada de la web 2.0 y la posterior creación de las redes sociales como Facebook o Twitter, el concepto de las redes sociales se ha vuelto más importantes para el análisis de la sociedad. En la actualidad, las redes sociales se consideran casi indispensables para muchas de las personas, ya que es la forma más sencilla de comunicarse con la gente de su entorno y para mantener el contacto en relaciones a distancia. Las redes también se han convertido en una herramienta indispensable para las empresas ya que con estas se tiene un trato más cercano con los clientes, convirtiéndose en un método más de contacto, ya sea para el área de atención al cliente o simplemente para publicidad de la compañía.

La creación de estas redes sociales está basada en la teoría de los Seis grados de separación, la cual dice que cualquier persona del mundo está conectada a través de no más de seis personas a cualquier otra persona del planeta [30]. Ésta se basa en la idea de que el número de personas que podemos alcanzar incrementa exponencialmente a medida que aumenta el número de enlaces. De esta teoría viene el potencial de publicitar una empresa a través de las redes sociales.



➢ ILUSTRACIÓN 1:EVOLUCIÓN DE USUARIOS EN LAS REDES SOCIALES[38]

En la ilustración anterior podemos ver como el número de usuarios activos en las redes sociales es cada vez mayor.

2.2.Importancia de las redes sociales.

Como bien se ha dicho anteriormente, la función principal de las redes sociales es la de comunicar a las personas y realizar nuevos vínculos de amistad ya sea a través de amigos o intereses comunes. Con el auge de las redes sociales y el impacto que estas han causado en la sociedad, haciendo que un gran porcentaje de la población tenga una cuenta en alguna red social, o incluso en varias, las empresas han visto aquí una manera de publicitarse y darse a conocer. Debido al gran auge que las redes sociales han sufrido, han surgido nuevos puestos de trabajo como el Social Media Manager o el Community Manager que ayudan a las empresas para saber cómo deberían comunicarse con sus posibles clientes, así como la publicidad que se les debería de enviar a estos a través de estos medios.

Existen numerosos tipos de redes sociales, aunque todos se basan en una misma idea, la relación entre las personas, ya sea de amistad o un interés

común. Para que cada una de estas redes se diferencie frente a las demás estas se han tenido que diferenciar mediante el contenido de las mismas. Este es el caso de LinkedIn, que permite a los usuarios crear sus currículum vitae de forma interactiva. Este es un gran ejemplo de lo importantes que se han vuelto las redes sociales en la actualidad donde estas no solo sirven para comunicarnos entre nosotros sino que ahora casi se hacen indispensables, incluso para conseguir un puesto de trabajo.

2.3.Historia de las redes sociales.

Una vez sabemos que es una red social y en que están basadas vamos a ver como estas han ido evolucionando a lo largo de la historia [3].

La primera red social que podemos considerar como tal es el e-mail, que permite la comunicación entre dos personas. El primer e-mail enviado entre dos ordenadores se realizo en el año 1971. Esto fue el fundamento de las bases para las futuras redes sociales, ya que como tal, fue la primera herramienta que se utilizo para la comunicación entre personas mediante internet.

En 1994 se funda GeoCities la primera red social de la historia. Esta se basaba en la creación de páginas sobre diversos temas, de tal forma que la gente interesada en un tema podía unirse y discutir [4]. Esta web estaba subdividida en seis barrios, de tal forma que cada barrio hacía referencia a una temática distinta, como por ejemplo los sitios relacionados con los ordenadores pertenecían a *Silicon Valley* mientras que los sitios dedicados a finanzas pertenecerían al de *Wall Street*. Desafortunadamente, pese a la gran popularidad que obtuvo en su momento, en el año 2009 se cerraron los servicios que Geocities ofrecía.

Viendo la gran popularidad que tomó Geocities, se desata el interés por las redes sociales creando redes tan populares actualmente como son Friendster (2002), Facebook (2004) y Twitter (2006).

Friendster fue la primera red social moderna que permitía intercambiar mensajes, fotografías y paginas entre los usuarios [5]. Además, permitía a los usuarios crear páginas con sus hobbies y la creación de eventos, por lo que era fácil el contacto entre personas con los mismos intereses. Debido a la aparición de nuevas redes sociales (Facebook), Friendster cambió su contenido de una red social pura a una basada en entretenimiento y videojuegos. Actualmente,

Friendster tiene su sede en Malasia ya que la mayor parte de su tráfico (90%) se sitúa en el sureste asiático.

La segunda gran red social de la que hablaremos es Facebook [6]. Esta fue diseñada por Mark Zuckerberg para los estudiantes de la universidad de Harvard. Debido al éxito que tuvo dentro de la universidad, esta se lanzó para todo el mundo, siendo actualmente la red social más utilizada con más de 1.440 millones de usuarios activos en marzo de 2015.

Entre las características de Facebook encontramos:

- Capacidad de agregar amigos.
- Chat que permite la interacción con otros usuarios de Facebook, previamente agregados, mediante mensajería instantánea.
- Publicación de notas en las paginas principales de tus amigos.
- Videollamadas usando el API-rest de Skype.
- Videojuegos propios dentro de su página.
- Centro de aplicaciones.

Es, sin duda alguna, la red social con más funcionalidad hasta el momento y es por ello sin duda que esta red social tiene una gran base de usuarios activos mensualmente.

Por último, Twitter es la red de microblogging más utilizada en la actualidad y es en la que estará centrada nuestra herramienta.

2.4.¿Qué es Twitter?

Twitter es un servicio de microblogging creado en marzo de 2006 por Jack Dorsey [7]. La red ha ido ganando popularidad mundialmente y se estima que tiene más de 560 millones de usuarios, solo por detrás de Facebook, generando 65 millones de mensajes al día y manejando más de 800.000 peticiones de búsqueda diarias. Actualmente es conocido por muchos como el "SMS de Internet".

La idea original de Twitter surgió dentro de la compañía Odeo, mientras se estaba desarrollando un servicio de radio on-line, también llamado podcasting, que no tuvo éxito debido al lanzamiento simultáneo de un producto similar por

parte de iTunes. Aun así, este prototipo de Twitter se utilizo durante un tiempo como forma de comunicación interna dentro de la propia empresa.

La versión definitiva se lanzó el 15 de julio de 2006. Su funcionamiento es muy sencillo, esta red se basa en el envió de una corta ráfaga de información limitada a 140 caracteres a los que se denomina "tweets".

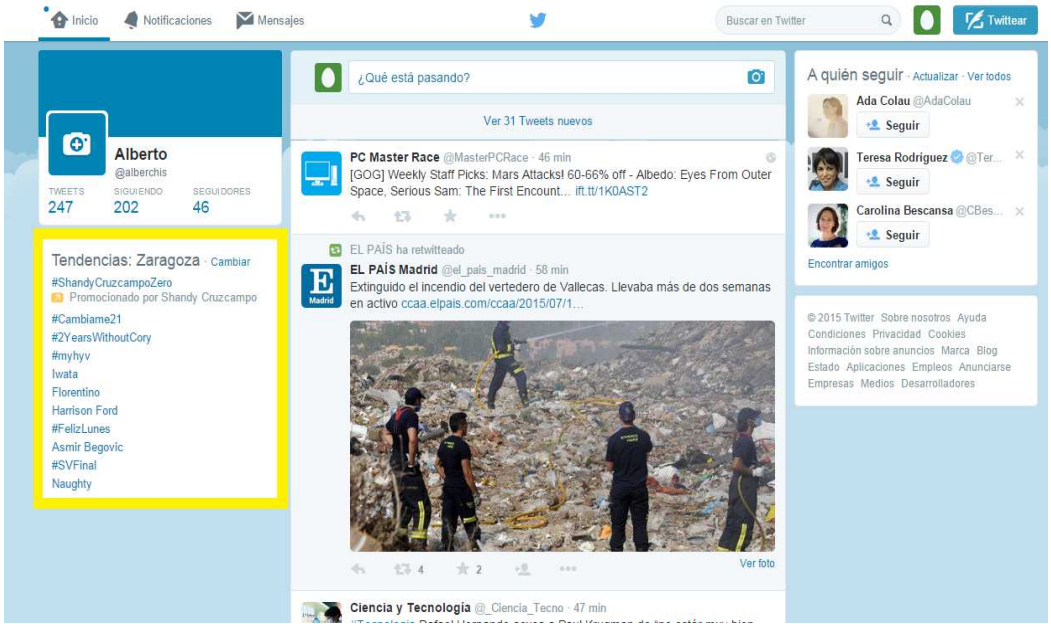
Twitter tiene una gran cantidad de términos propios como son:

- **Hashtag:** El hashtag son palabras que se utilizan para realizar una tendencia. Los hashtag son precedidos por el símbolo # junto a la palabra que se desea hacer visible. Los mensajes que incluyan un hashtag tienen más impacto dentro de la red debido ya que este símbolo se usa a modo de etiqueta y nos permite buscar mensajes con dicho hashtag más fácilmente. De esta forma es más probable que una palabra precedida por este símbolo acabe siendo Trending Topic [31].

- **Trending Topic o TT:** Los Trending Topic son una de las características más relevantes en Twitter ya que ninguna otra red social se centra tanto en este aspecto. Los Trending Topic son las palabras más utilizadas dentro de una zona geográfica determinada. Es una de las herramientas más potentes que nos ofrece Twitter ya que nos permite conocer cuáles son las tendencias en cada zona del mundo [32].

  Estas zonas geográficas se determinan mediante un código de geolocalización, que permite diferenciar cada zona geográfica. Para determinar estos Trending Topic, Twitter utiliza un algoritmo de tal forma que los Trending Topic se ajustan a cada uno de los usuarios en función de la gente a la que siga.

  Esta herramienta nos permite saber con un simple vistazo cuales son los temas de cada zona, ya que esta lista de Trending Topic aparece en la página principal de cada usuario como se muestra en la siguiente ilustración.



➢ ILUSTRACIÓN 2:TRENDING TOPIC.

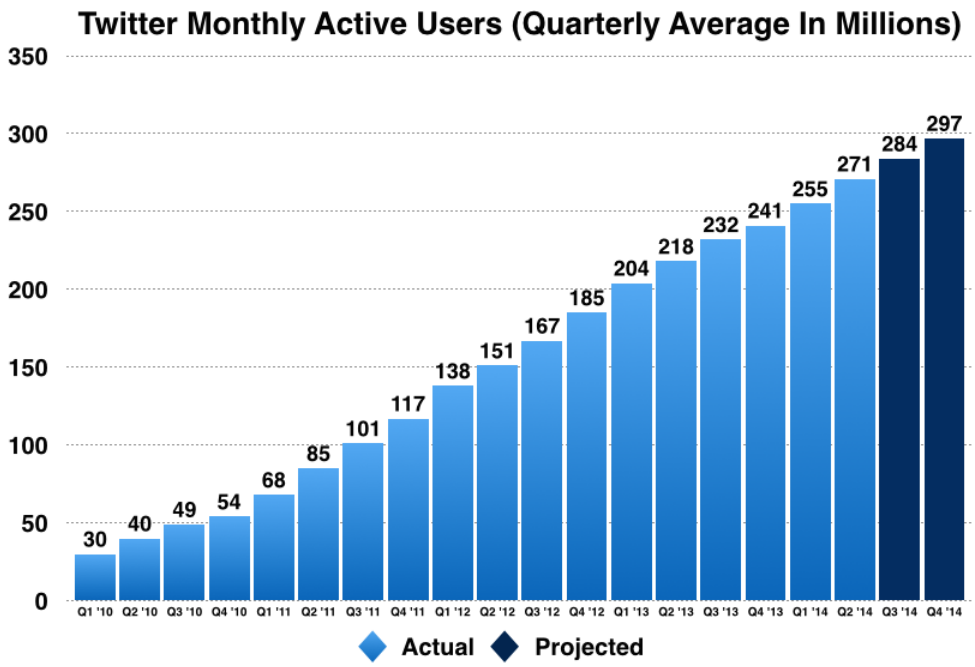
Como vemos en la ilustración, los Trending Topic más relevantes del momento aparecen en la parte izquierda inferior de la página principal del usuario por lo que es una información que esta siempre disponible para el usuario.

- **Promoted Tweets:** En abril de 2010 Twitter lanza una nueva estrategia para emitir publicidad en forma de un servicio llamado Promoted Tweets, que podría traducirse como "tuits patrocinados". Esta idea se basa en la creación de tuits con el patrocinio de alguna empresa, apareciendo estos tuits como primer resultado cuando miremos nuestro TimeLine, que es la página donde se muestran los mensajes de las personas a las que seguimos. El número de estos Promoted Tweets está limitado a uno para evitar la saturación de publicidad en los usuarios. Estos tuits suelen ser dirigidos, es decir, según a la gente que sigas el tuit que recibirás de publicidad será distinto, adecuándose a los intereses de los usuarios. Políticas parecidas a estas han sido desarrolladas por Google o Facebook para ofrecer publicidad personalizada y maximizar el interés por el anuncio. Este servicio de publicidad está disponible para los desarrolladores de aplicaciones de forma que puedan obtener beneficios de esta.

- **Promoted Trending Topic:** Al igual que ocurre con los Promoted Tweets, el propio Twitter ahora ofrece la oportunidad a las empresas de tener un hashtag entre los Trending Topic del momento. De esta forma las empresas al contratar estos servicios son capaces de llegar a un número más elevado de usuarios [12].

Con estos dos últimos elementos que hemos descritos se ve la importancia que está tomando dentro de Twitter la publicidad tanto con los mensajes publicitarios personalizados como los Trending Topics.

Además, con el gran número de usuarios con los que cuenta, Twitter se ha convertido en la segunda red social más importante, solo por detrás de Facebook y se ha vuelto indispensable para la transmisión de información en directo, debido a su limitación de caracteres, ya que permite a los usuarios, en un simple vistazo, ver las noticias más importantes del día con los temas más hablados del momento o también conocidos como Trending Topic.



➢ ILUSTRACIÓN 3:EVOLUCIÓN USUARIOS TWITTER[8]

Como vemos en la imagen anterior, Twitter se encuentra en constante crecimiento. Es por ello que se muestra como una de las redes sociales con un futuro más prometedor.

Al igual que ha hecho Facebook, el número de herramientas vinculadas a Twitter ha ido creciendo y ahora no solo podemos mandar mensajes de texto planos, sino también enlaces a páginas, imágenes e incluso videos dentro de los propios mensajes.

3. Spam

Desde su origen, el principal objetivo del spam ha sido la monetización, ya sea mediante la venta de un producto o mediante la publicidad de una empresa. En nuestro caso, utilizaremos el spam en nuestra herramienta para ganar relevancia dentro de una red social.

Dentro de esta sección analizaremos el spam dentro de las redes sociales y en especial Twitter, de tal forma que nos haremos una idea de las malas conductas que debemos evitar y las soluciones que debemos implementar para que la propia red no nos bloquee el acceso.

3.1.¿Qué es el spam?

Se denomina spam, correo basura o mensaje basura a los mensajes no solicitados, no deseados o sin remitente que perjudican de alguna manera al receptor del mensaje. El contenido de estos mensajes suele ser publicitario y se suele enviar a un gran número de personas con el fin de tener la máxima repercusión posible [10].

El primer mensaje de spam fue enviado a través de correo electrónico en el año 1994. El mensaje fue generado por una firma de abogados, que envió este mensaje a una lista de e-mails de amigos y lectores de la red donde se daba a conocer la empresa. Debido al éxito de facturación que tuvieron en los días siguientes, vieron que esta forma de publicidad era posible. Desde entonces, el número de empresas que utilizan este método de publicidad ha ido creciendo a lo largo de los años.

Aunque actualmente, debido a los progresos en la informática, estos mensajes se pueden enviar a través de cualquier vía de conversación, donde más se utiliza es a través de correo electrónico.

3.2.Impacto económico

Cuando hablamos del impacto económico del spam se debe hablar del coste que este supone a las empresas lidiar con este problema, así como de los beneficios que se generan debido a las ventas por estos mensajes no deseados [11]. Desde este punto de vista, el dinero gastado (esto incluye: el desarrollo de herramientas de antispam, coste computacional para el análisis del correo, uso

de red...) para prevenir los mensajes no deseados es 100 veces mayor que las ganancias que se generan debido a estos.

Actualmente se estima que de los 50000 millones de mensajes de spam enviados diariamente, entre el 1.8% y el 3% de estos llegan a su destinatario. Es por esto, que el coste computacional que se necesita en la detección de estos correos no es nada despreciable.

Se estima que para que un correo de spam sea efectivo, este debería vender un produzco cada 25000 mensajes recibidos, por lo que a pesar de los programas antispam resulta una publicidad efectiva para las empresas.

Además de los gastos previamente comentados hay que indicar que los correos que logran llegar a su destinatario, si se tratan de correos laborales el tiempo invertido en la lectura y eliminación de estos correos también supone un gasto, debido a la reducción del tiempo efectivo de trabajo.

Es por ello que, a pesar de que en sus orígenes este tipo de publicidad si era efectiva, actualmente ha perdido poder en los usuarios. Además, el uso de estas herramientas suponen grandes pérdidas en las empresas.

3.3.Correo masivo en diferentes medios

Como bien hemos comentado anteriormente, el spam se utiliza a través de cualquier herramienta que permita la comunicación de usuarios. Es por esto que la creación de estos mensajes basuras se ha ido especializando en cada uno de los ámbitos, permitiendo de esta forma que la información del mensaje se adecúe más al contexto y que los mensajes superen los distintos filtros que hay en cada una de las redes de comunicación.

A continuación, hablaremos un poco sobre los distintos medios en lo que se envían los mensajes no deseados.

Spam en el correo electrónico.

Además de ser el primer medio de comunicación que soportó los primeros mensajes de spam también es en la actualidad el medio que soporta la mayor parte del tráfico producido por spammers. Estos suelen ser utilizados para dar a conocer empresas u ofertas publicitarias de algún producto específico.

Normalmente los correos generados como spam no tienen remitente o esta dirección es falsa, por lo que la respuesta a estos suele ser inútil. Pesé a que no se puede verificar si una dirección remitente es real o falsa, lo cual dificulta la tarea de localizar posibles fuentes de spam, se puede firmar los mensajes escritos con la clave pública del emisor del mensaje que permitan facilitar esta labor.

Actualmente los filtros automáticos anti-spam se basan en la búsqueda de palabras clave como *rolex*, *viagra* y *sex*, que puedan connotar que se trata de un mensaje de publicidad no deseado por el usuario. Es por ello que se deben evitar el uso de estas palabras en correos para evitar el riesgo de que sean catalogados como correos no deseados.

Spam aplicados en los blogs.

Con el auge de los blogs, los spammers, o gente que se dedica hacer spam, han visto un nicho donde la emisión de mensajes puede resultar beneficioso. Este tipo de spam se basa en la creación de un mensaje que se postea en los comentarios, vinculado con un link a la pagina que se desea dar publicidad. Actualmente no hay muchas herramientas que eviten este tipo de mensajes, aunque WordPress, empresa facilitadora de blogs, ha creado un complemento que permite detectar los correos masivos en su red.

Spam en las redes sociales.

Poco a poco, las redes sociales se ven sometidas a mas contenido viral y es debido principalmente a la gran popularidad de la que gozan. Dependiendo de la red social, cada una tiene unas características u otras. Entre las técnicas más utilizadas de spam dentro de estas redes se encuentra la publicidad directa mediante mensajes privados aunque también existen otros métodos más característicos de cada red social, como por ejemplo Twitter permite el mensaje de Trending Topics o Facebook, que mediante los muros de cada usuario permite una gran visibilidad de los comentarios.

Esta se está convirtiendo en una de las formas de spam predominante en los últimos años. Normalmente, debido a las políticas anti-spam de las redes sociales sus cuentas suelen ser eliminadas a las pocas horas de su creación. Es por ello que las técnicas utilizadas para este tipo de spam suelen ser más complicadas con el fin de no ser detectadas fácilmente.

Además de las técnicas anti-spam propias de las redes sociales, muchas de las redes sociales ofrecen a los usuarios la capacidad de denunciar a otro usuario, ya sea por contenido inapropiado o ya sea por contenido de spam. Esto hace que, los spammers en las redes sociales tengan, generalmente, un tiempo de vida más corto que en otro medios de comunicación.

Esta técnica de spam es la que tiene una mayor relación con nuestro proyecto, ya que debido a que nuestra herramienta va a estar publicando grandes cantidades de mensajes con varios usuarios virtuales se podría considerar que estamos realizando spam dentro de la red social.

Spam en foros.

El spam en los foros se define como aquellos mensajes que no tienen que ver con el tema expuesto o que no contribuyen de forma alguna al desarrollo de la conversación. A este caso hay que unir los previamente comentados en lo blogs, donde un usuario escribe un mensaje con un link vinculado al producto o empresa que se desea publicitar.

Recientemente, algunos foros han creado hilos alternativos donde solo se publican mensajes de spam publicitario. Estos han tenido un gran éxito, siendo en muchos casos los hilos más activos del foro.

Spam en redes de IRC

Debido a la masificación de los chats IRC el uso de spam se ha visto promocionado. Esto unido al coste tan bajo que supone la emisión de un mensaje en este tipo de plataformas han hecho que estos chats sean un gran foco de exposición a los mensajes no deseados. Estos mensajes al igual que en otros medios lo que buscan es la publicidad de alguna página o producto. Además, en estos chats IRC también ha surgido un nuevo método de spam basado en la petición del número de telefonía móvil para contratar servicios de pago.

Spam en correo postal y vía pública.

Las técnicas de spam no solo están presentes en los medios a través de internet sino que también están en el correo convencional. Debido a los bajos costes que supone la impresión de panfletos, incluso nuestros buzones de correo físicos sucumben a los mensajes publicitarios no deseados. Como ocurre con los e-

mails, estos mensajes publicitarios carecen de dirección de origen, pero el spam sigue llegando a su destinatario.

Al igual que ocurre con el spam en el correo postal, esta es otra de las practicas de spam mas utilizadas, donde se entregan panfletos a la gente en lugares estratégicos, con el fin de que conozcan tu producto.

3.4.Técnicas de spam.

A lo largo de los años, las técnicas en el filtrado de mensajes no deseados han ido mejorando poco a poco, haciendo que los spammers tengan que ser cada vez más ingeniosos para llegar a los clientes potenciales. De esta forma, las técnicas de spam han ido evolucionando con los programas antispam con el fin de mejorar el índice de mensajes que llega a los usuarios [13,14]. Dentro de las mejoras de los mensajes de spam existen dos grandes áreas para la mejora de la generación de spam: la evasión de la dirección de remitente y el contenido de los mensajes.

Técnicas de emisión.

A continuación se va a hablar sobre la evolución que han ido sufriendo los mensajes en cuanto a las técnicas de emisión con el fin de que este pasará los filtros anti-spam.

- Al principio debido a la inexistencia de filtros antispam era sencillo el introducir estos mensajes en las bandejas de entrada de los usuarios, por lo que se enviaban directamente sin importar la dirección de emisión.

- Una vez se introdujeron las IP "tóxicas" en los filtros antispam surgieron los servidores de correo abierto, que permitían la emisión de cualquier mensaje a cualquier destinatario.

- Cuando este método se volvió más ineficiente se explotó la debilidad de los ISP, que creaban direcciones IP dinámicas para los usuarios, de esta forma podían hacer spam con diferentes IP desde una única máquina de envío.

- Los ISP entonces regularon el número máximo de e-mails que se podían enviar con cada sesión, limitando así el potencial de la técnica anterior descrita. Es por ello que con la llegada del nuevo siglo se empezaron a usar proxys para la emisión de los mensajes haciendo que los mensajes utilizaran la IP del proxy en cuestión.

- Actualmente, junto a los proxys, lo más utilizado es el hackeo de maquinas de usuario mediante troyanos descargados a través de redes P2P o gusanos a través de e-mail. De esta forma, cada maquina hackeada permitirá al spammer enviar correos desde una IP distinta.

Técnicas de contenido.

Al igual que con las técnicas de emisión el contenido de los mensajes de spam ha ido evolucionando a los largo de los años debido al avance de los sistemas antispam, con el fin de que este no sea descubierto. Las técnicas que se han ido utilizando son:

- Incrustación de texto y contenido HTML.
- Mensajes personales.
- Texto aleatorio y texto invisible. Esto permitía evadir el análisis estadístico, ya que el contenido de los diferentes mensajes de spam era distinto.
- Envío de imágenes. Al enviar imágenes en lugar de texto plano se impedía que se viera la información de una forma clara.
- Textos parafraseados. Debido a la reestructuración de la información dentro del mensaje y utilizar palabras distintas se evita la detección de los correos no deseados.

Actualmente las técnicas más utilizadas son las tres últimas, ya que son las más efectivas con los programas antispam actuales debido a la complejidad que conlleva su detección.

3.5. Spam en las redes sociales

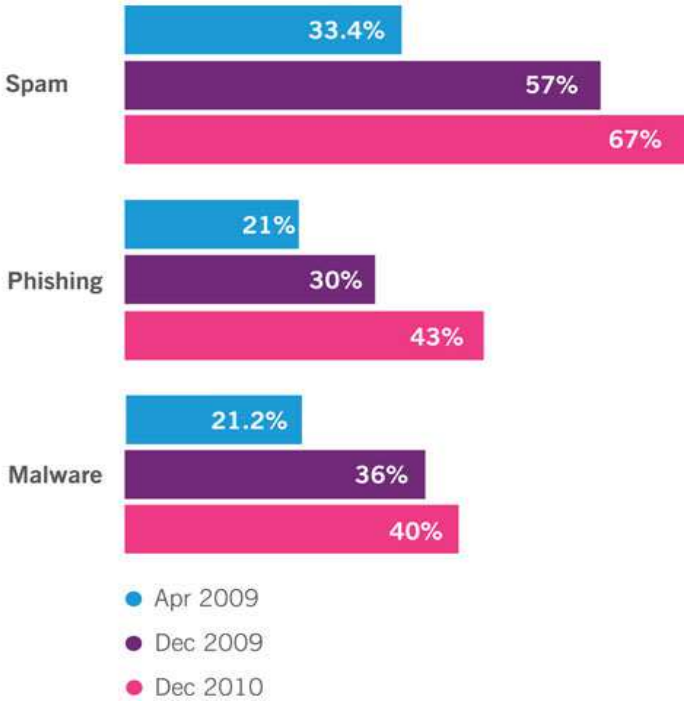
Debido al auge de las redes sociales, la llegada del spam era cuestión de tiempo. Las redes sociales más importantes, como Facebook o Twitter, se ven sometidas a una gran cantidad de spam día a día. Se calcula que 5 de cada 7 cuentas creadas son de spammers y que el tráfico que suponen es alrededor de 1 de cada 200 mensajes enviados a la red son de spam [33].

La erradicación de las cuentas de spam es importante en estos medios, ya que una exposición prolongada a contenido no deseado generaría la desconfianza de los usuarios, perdiendo de esta forma potencial en la red. Es por ello que

últimamente se está destinando muchos recursos con el fin de acabar con este problema del spam.

Las redes sociales se llenan mas de spammers día a día. Esto es debido principalmente a dos cosas: el auge de las redes sociales en estos últimos años, donde casi todo el mundo tiene una cuenta en alguna red, y la facilidad de comunicación, ya que en estas se permite la comunicación a un gran número de personas cosa que el correo electrónico no permite con facilidad.

Aun así, en la siguiente imagen vemos como el número de mensajes tanto de spam como de malware, han aumentado en los últimos años en todas las redes sociales.



➢ ILUSTRACIÓN 4:SPAM Y MALWARE EN LAS REDES SOCIALES[39].

En la imagen se muestran el porcentaje de los usuarios encuestados que ha recibido mensajes de spam, phishing o malware. Como vemos, cada vez un mayor número de usuarios se ve afectado por este fenómeno en todas las redes sociales.

Las técnicas utilizadas son muy variadas y van desde el desarrollo de aplicaciones que solo contienen publicidad hasta las cuentas falsas, pasando por los bots capaces de escribir por si solos.

Spam en Twitter

Actualmente el spam en Twitter aún es relativamente nuevo y aún no alcanzado su auge como el del correo electrónico. Aún así, poco a poco, el número de cuentas creadas únicamente para realizar las tareas de spam está creciendo [16,17]. Es por ello que Twitter ha aumentado sus esfuerzos en la detección de estas cuentas con el fin de preservar la integridad de los usuarios de la plataforma.

A continuación hablaremos de las técnicas que se utilizan principalmente para generar el spam en Twitter, las metas de este spam así como las técnicas que utiliza el propio Twitter para identificarlas.

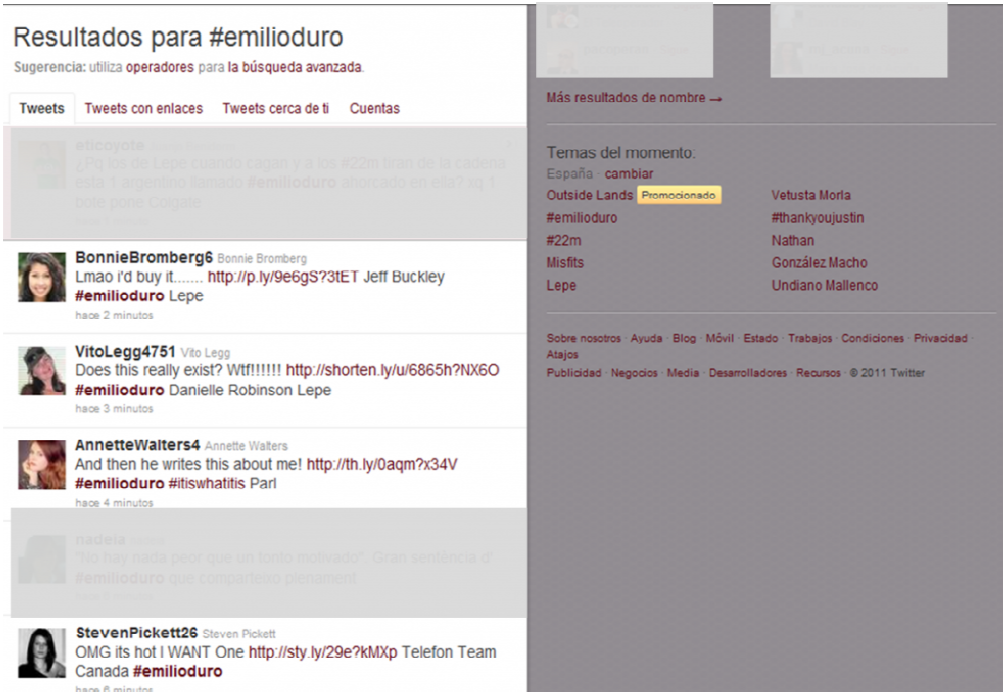
Generación de spam en Twitter.

En general, existen cuatro formas distintas por las que la gente se suele ver expuestas a este tipo de mensajes no deseados en Twitter [18]. Estas son:

- **TimeLine del usuario:** En esta técnica los mensajes llegan a la página principal de los usuarios. Para ello los spammers se valen de dos herramientas, o los usuarios son mencionados dentro del tweet , o bien el usuario sigue al spammer. Pese a que la segunda es un poco más complicada de realizar en cuentas dedicadas únicamente a spam, la primera es sencilla, ya que Twitter permite realizar estas menciones sin necesidad de conocer a la otra persona. Aun así, mediante el hackeo o phishing de las cuentas de Twitter se puede conseguir que las cuentas de los usuarios sigan a los spammers.

- **Trending Topics:** En este método de emisión el spammer utiliza los Trending Topics del momento para dar publicidad a su mensaje. Debido a la herramienta que ofrece la compañía para ver los temas más hablados del momento, esto hace de este método una manera efectiva de que el spam llegue a un gran número de personas.

  En la siguiente imagen vemos como los usuarios que realizan el spam aprovechan uno de los temas del momento para realizar publicidad sobre sus páginas web.



➢ ILUSTRACIÓN 5:SPAM CON TRENDING TOPICS[35]

- **Búsqueda:** Esta técnica se basa en la creación de los mensajes con palabras que la gente busque, de esta forma para palabras muy usadas el spam llegará a más personas. Aun así, esta forma de spam es peor que las demás ya que la técnica de búsqueda no es tan utilizada como la de los Trending Topics.

- **Mensajes directos:** En este último método se da uso a los mensajes directos que proporciona Twitter para mantener conversaciones privadas. Esta técnica sería muy parecida a la que se utilizaría para correo electrónico, ya que cada mensaje enviado llegaría a un único destinatario. Aun así, este es un método más eficiente que por correo electrónico, ya que mediante esta herramienta no tendrías que saltar los programas antispam. Sin embargo, debido a las últimas mejoras realizadas por Twitter, el envío de mensajes privados solo se puede realizar a aquellas personas que te siguen, por lo que el uso de esta técnica es un poco más complicado de realizar de lo que lo era previamente.

Objetivos del spam en Twitter.

El objetivo principal del spam es la monetización y en Twitter no iba a ser menos. Con el fin de cumplirlo existen tres formas distintas de obtener dinero mediante esta red social:

- **Programas afiliados:** En este método se utilizan unas webs de venta de productos (generalmente clickbank y amazon) que mediante la compra a través de un link especifico se obtiene una remuneración económica. Por lo tanto, en este caso los tweets lo que intentan es vender productos en estas webs a través del enlace que estas les proporcionan.

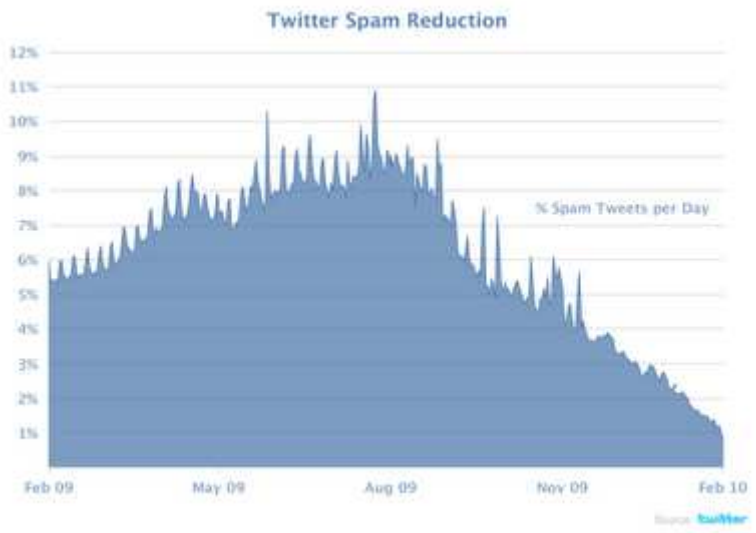
- **Vendedores de cuentas:** En esta segunda opción se usan programas de phishing con el fin de poder suplantar las cuentas de usuarios. Normalmente se mandan enlaces a webs capaces de realizar esta tarea para posteriormente vender el uso de estas cuentas, generalmente a otros spammers.

- **Publicidad:** En esta última se utilizaran los conocidos recortadores de URL para la emisión de publicidad a través de esta.

Contramedidas de Twitter.

En esta última sección hablaremos de las contramedidas que ha tomado Twitter para evitar la proliferación de cuentas de spammers en su red.

Las principales medidas que Twitter ha tomado se basan en la creación de una lista negra (tanto de generación automática como manual) que permite identificar muchas de las cuentas nada mas estas son creadas. Además de esta lista negra también se procede al análisis de los tweets , sobre todo el de nuevas cuentas, que en caso de tener contenido de spam (esto puede ser cualquiera de los tres casos comentados en el punto anterior) se procedería a su cancelación.

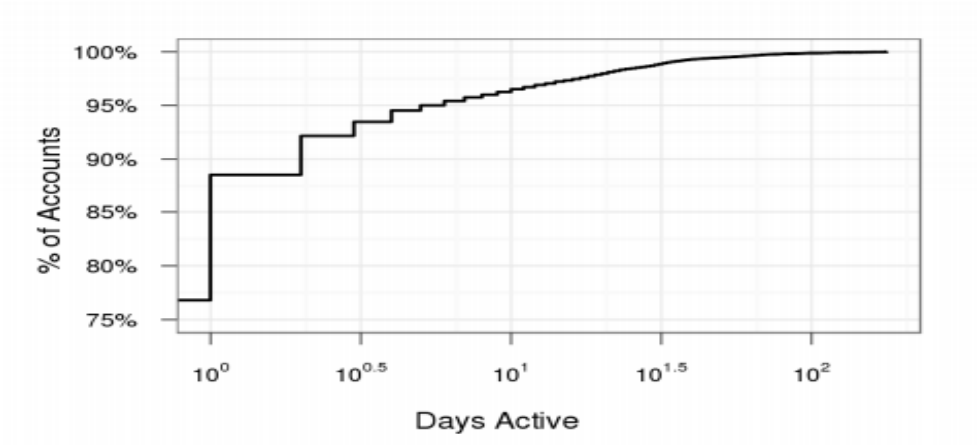


➢ ILUSTRACIÓN 6:SPAM EN TWITTER

Como vemos en la imagen anterior estos métodos parecen funcionar contra los spammers aunque se debe seguir mejorando esta contramedida, ya que como se dijo al principio, aun es una forma de spam nueva, por lo que cada vez estos mensajes no deseados se volverán más y más difíciles de identificar y detectar [34].

Para mejorar el sistema, además de la lista negra y el análisis de los tweets, las redes sociales proveen a los usuarios la capacidad de denunciar a las cuentas que sean perjudiciales, ya sea por contenido inapropiado o ya sea por contenido de spam. Esto hace que, los spammers en las redes sociales tengan, generalmente, un tiempo de vida más corto que en otro medios de comunicación.

Normalmente, debido a las políticas anti-spam de las redes sociales sus cuentas suelen ser eliminadas a las pocas horas de su creación. Es por ello que las técnicas utilizadas para este tipo de spam suelen ser más complicadas con el fin de no ser detectadas fácilmente.



➢ ILUSTRACIÓN 7: TIEMPO DE VIDA DE LAS CUENTAS DE SPAM[18].

Como vemos en la ilustración la mayoría de las cuentas de spam que se abren en esta red social son detectadas nada más son creadas, cuando inician la emisión de mensajes. Es por ello que aún los spammers necesitan especializarse más en estos medios de comunicación.

Además de las contramedidas previamente comentadas, Twitter tiene una serie de normas para los usuarios con el fin de que se respete los derechos de los demás [19]. La propia web de Twitter las divide en dos grupos, usos y limitaciones y abuso y spam.

En la primera, las normas involucradas tienen que ver con el uso que los usuarios dan a la propia red de Twitter, de tal forma que no se realicen cosas ilegales dentro de la propia red.

En la segunda se regulan el abuso que los usuarios pueden provocar a terceros y la influencia de las cuentas de spam.

En nuestro caso, vamos a analizar los puntos que Twitter considera como perjudiciales en lo que spam se refiere.

Twitter y el spam

A continuación mostramos la información a partir de la cual Twitter puede considerar nuestra cuenta como cuenta spammer y bloquearnos el acceso a la red. Debemos saber cuál es la información que utiliza Twitter para declarar una cuenta como spam ya que nuestra herramienta va a utilizar un gran número de cuentas virtuales para publicar automáticamente contenido. Si conocemos cuales son las verificaciones que realiza Twitter para catalogar una cuenta como

spammer lo que haremos será evitar esa clase de escenarios, de tal forma que nuestra cuenta sea más difícil de detectar y tenga un periodo de vida útil más largo.

Sin más, vamos a ver cuáles son los datos que Twitter tiene en consideración para etiquetar una cuenta como spammer.

Spam: el usuario no podrá utilizar el servicio de Twitter con el fin de enviar spam a nadie. Algunos de los principales factores que Twitter tendrá en cuenta para determinar qué conducta se considera envío de spam son:

- si el usuario se hace seguidor y/o deja de seguir a grandes cantidades de usuarios en un corto periodo de tiempo, especialmente por medios automatizados (también denominado como seguimiento agresivo o seguimiento intermitente);
- si el usuario sigue y deja de seguir de manera repetida, ya sea para conseguir seguidores o para atraer más atención al perfil;
- si las actualizaciones son principalmente enlaces y no actualizaciones personales;
- si un gran número de personas ha bloqueado al usuario;
- si se ha recibido un gran número de quejas por spam en contra del usuario;
- si el usuario publica contenido duplicado en múltiples cuentas o múltiples actualizaciones duplicadas en una cuenta;
- si el usuario publica múltiples actualizaciones sin relación con un tema mediante #, un tema popular o del momento o una tendencia promocionada;
- si el usuario envía un gran número de @respuestas o menciones duplicadas;
- si el usuario envía un gran número de @respuestas o menciones no solicitadas;
- si el usuario agrega un gran número de usuarios sin relación a listas;
- si el usuario crea reiteradamente contenido falso o engañoso;

- si el usuario sigue, marca como favorito o retwittea Tweets de forma aleatoria o agresiva;

- si el usuario publica de manera repetida la información de la cuenta de otros usuarios como propia (biografía, Tweets, url, etc.);

- si el usuario publica enlaces engañosos (por ej., enlaces a afiliados, enlaces a malware/páginas de clickjacking, etc.);

- si el usuario crea cuentas o interacciones de cuenta engañosas;

- si el usuario compra o vende interacciones de cuenta (es decir, compra o vende seguidores de Twitter, Retweets, favoritos, etc.);

- si el usuario usa o promociona servicios o aplicaciones de terceros que afirman conseguir más seguidores (por ejemplo, trenes de seguidores, sitios que prometen "más seguidores con rapidez" o cualquier otro sitio que ofrezca agregar seguidores automáticamente a su cuenta).

Dentro de las reglas, las que más nos van a afectar en nuestro proyecto son:

- creación reiterada de contenido falso o engañoso.

- marcado como favorito o retwittea Tweets de forma aleatoria o agresiva.

- creación de cuentas o interacciones engañosas.

Estas serán las normas que restringirán en mayor medida nuestro grado de actuación en cada uno de los usuarios virtuales.

Para la primera norma utilizaremos un generador de lenguaje, que nos permitirá generar contenido aleatorio para cada uno de los usuarios, de tal forma que, aunque no todas las frases tengan sentido, las cuentas publicaran contenido diferente cada vez que envíen un mensaje a la red.

En cuanto a las otras dos normas estas nos limitan la capacidad de actuación e interacción entre las cuentas de spam, por lo que limitará el número de interacciones que realizarán las cuentas.

Ejemplos de utilización con bots.

Uno de los ejemplos de utilización de spam en Twitter fue la creación del bot @Trackgirl [20]. Este bot, creado por Greg Marra, escribía tweets diarios sobre las supuestas experiencias que vivía una chica atlética. Lo interesante del proyecto es que la gente se acabo involucrando con ella llegando a preguntarla que tal iba mejorando tras su supuesta lesión. Pese a que este bot fue diseñado para ver como se distribuía la red de amigos dentro de una comunidad, en este caso tracking, se puede ver como a través de la creación de un bot se puede llegar a influir dentro de una comunidad en Twitter. Además, debido a la estructura de mensajes limitados a 140 caracteres que posee Twitter y que estos no suelen tener coherencia entre ellos, hicieron que este bot fuera mucho más creíble que pudiera haber sido en otra red social.

4. Generación de lenguaje(NLG)

4.1.¿Qué es la generación de lenguaje?

La generación natural de lenguaje o como es conocido en inglés, Natural Language Generation (NLG) es una rama de la inteligencia artificial y la lingüística computacional que se encarga de la construcción de los sistemas informáticos capaces de generar texto comprensible para el ser humano en cualquier lenguaje[21].

Como hemos dicho en el apartado anterior, el generador de lenguaje se encargará de generar distintas frases para cada uno de los usuarios virtuales que estén activos. De esta forma, estos no llegarán a ser detectados como spam y podrán seguir publicando mensajes.

La tarea principal de un generador de lenguaje es la traducción o mapeado de unos valores de entrada a unos valores de salida. Esta tarea de traducción se puede dividir en pequeñas tareas de tal forma que el sistema se haga más modular simplificando las tareas de cada uno de estos subsistemas. Estas pequeñas tareas son seis.

Elección del contenido:

Una de las principales tareas de los sistemas de generación de lenguaje es la elección de las palabras que se desean utilizar. Esto es una de las tareas principales ya que sin una elección apropiada de estas palabras la frase puede significar otra cosa totalmente distinta. En esta primera fase se generan varios mensajes ya sea con la información de entrada que nos proporcionen o mediante la base de datos de palabras. El conjunto de palabras elegido será el que posteriormente utilicen las siguientes tareas. Tanto el contenido como la forma de presentación de los mensajes depende del escrito que deseemos realizar. Es por ello que la función básica de este sistema se basa en el filtrado de las palabras que se desean utilizar, y una vez seleccionadas en la representación formal de estas para que las futuras tareas puedan acomodar los mensajes al tipo de texto que se desee generar. Esta representación formal de los datos se hace mediante el uso de entidades, conceptos y relaciones entre las palabras que permitirán a las futuras tareas realizar una estructuración del contenido de una forma más eficiente.

Aquí solo se escogen las palabras principales que se deseen utilizar en el mensaje, es decir, solo las que aportan el mayor significado al texto.

Planificación del discurso:

Esta tarea se encarga de la organización del escrito que se desea generar. Es decir, su función es, en función de los datos de entrada provistos por el elector de contenido, este decide el orden de cada una de las frases con el fin de que este sea lo más legible posible. La salida de este sistema suele consistir en una estructura de tipo árbol, en la cual cada una de las hojas muestra uno de los mensajes previamente seleccionados por el elector de contenido y los nodos intermedios representan la relación que tienen entre ellos. El orden que tomen en el árbol cada uno de los mensajes es importante ya que indica el orden en el texto final.

Unión de sentencias:

En este sistema se unen las distintas oraciones previamente ordenadas. Esto no solo se refiere a la concatenación de las distintas frases mediante frases independientes sino que, si se aplica correctamente, poder enlazar distintas frases con conectores. Este sistema es importante para mostrar una sensación de fluidez cuando se genere el texto final. Además, si el sistema está bien codificado, este podría separar correctamente la información entre párrafos permitiendo que en cada uno de ellos se tratará uno de los temas de los que consta el texto.

Lexicalización:

Se define como el sistema que elige las palabras para expresar lo que desea escribir. Este sistema se basa en la información adicional que proporcionaba el sistema de elección de contenido que decía la relación que tenían cada una de las palabras. En este sistema se decide, por ejemplo, los verbos que se utilizarán en cada una de las frases, en función del significado que deseamos proporcionar puede que muchos verbos se ajusten al mensaje pero alguno se acopla mejor al contenido del mensaje.

Este sistema es muy importante sobre todo cuando se desea realizar texto en distintos idioma ya que para cada uno de ellos la palabras adecuadas para el texto varía y la traducción literal del contenido puede no expresar la información correctamente.

Generación de expresión:

Es la parte encargada de realizar la expresión del texto, es decir, que el texto tenga sentido entre frases. La tarea de este sistema está muy ligada al anterior, el lexicalizador, ya que la tarea de ambos es muy parecida. Sin embargo, este se basa en la discriminación de contenido donde lo que se trata hacer es la distinción entre dos dominios del texto. Para ello, generalmente se basa en la información generada previamente, el contexto, para saber si la información que se debe escribir se puede reescribir de otra forma. Con esto conseguimos que el texto generado sea más parecido al que haya podido generar un humano.

Realización lingüística:

Por último, el realizador lingüístico se encarga de aplicar las normas de gramática: morfología, sintaxis y ortografía. Este sistema se suele basar en un diccionario y unas normas que, dependiendo de la frase que se le pase por parámetro aplicara de una forma u otra.

4.2.Arquitecturas para el generador de lenguaje

Una vez se han visto las tareas que debe realizar un sistema generador de lenguaje podemos ver los diferentes tipos de arquitecturas.

Lo más sencillo sería la realización modular de una arquitectura, donde cada uno de los módulos de la misma cumpliera una de las tareas previamente explicadas. Esta sería una buena solución ya que nos permitiría modificar cada uno de los módulos por separado, de tal forma que si uno de estos no funcionara como se esperaba, se podría intercambiar o modificar de una forma sencilla, sin interferir en los demás sistemas.

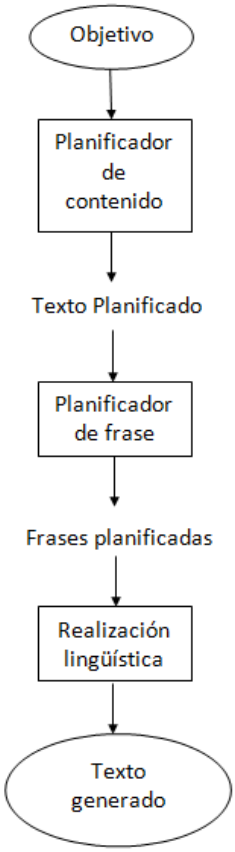
Sin embargo, la arquitectura más utilizada en la actualidad no es esta, y es una basada en tres etapas. Las etapas serían:

- **Planificación de contenido:** En esta primer etapa se realizarán las dos primeras tareas que debe realizar un NLG, la elección del contenido y la planificación del discurso.

- **Planificador de frases:** En esta fase se realizarán las tres siguientes tareas: unión de sentencias, lexicalización y la generación de expresión.

- **Realización lingüística:** Por último, en esta etapa se realizará la tarea con el mismo nombre.

De tal forma que el diagrama del sistema sería el siguiente:



➢ ILUSTRACIÓN 8:DIAGRAMA DE NLG

Como hemos dicho previamente, este es uno de los diagramas de sistemas más utilizado a la hora de realizar un Generador Natural de Lenguaje y es debido a que cada uno de los módulos de los que consta el sistema tiene una tarea concreta que realizar.

El primero de los módulos se encargara de decidir cuál es la información que se quiere incluir dentro del texto que se va a generar. Es por ello que este primer modulo realiza las dos primeras tareas, ya que se encargan tanto de la selección de las palabras que se van a incluir así como de la relación que van a tener entre ellas.

En el segundo módulo se realizará la correcta sintetización y unión de las frases. La tarea principal por tanto es la de generar un texto más compacto y con una mayor riqueza lingüística posible, de tal forma que el texto generado sea más parecido al que pueda crear un ser humano.

En el último modulo, por tanto, solo se realizará la función sintáctica, es decir, solo se corregirán los posibles errores que se hayan podido producir a la hora de ir introduciendo o sintetizando contenido dentro del texto.

Por último, a la hora de realizar sistemas basados en distintos módulos es necesario definir un sistema de entrada y salida de datos, de tal forma que los módulos puedan entender la información que se genera en los sistemas anteriores. Esto se debe definir a la vez que se definen el número de módulos de los que constará el sistema y la función que desempeñará cada uno de estos.

SimpleNLG

Antes de pasar al siguiente apartado vamos a comentar una librería que nos permite generar lenguaje. Este es el caso de SimpleNLG, que es una librería basada en Java, totalmente gratuita, que nos permite generar frases en inglés de una forma muy sencilla [22].

Haciendo referencia al apartado anterior donde se hablaba sobre las arquitecturas posibles y cada uno de los módulos, esta librería nos ofrece una corrección sintáctica de la frase generada, de tal forma que lo único que debemos de preocuparnos es acerca de la información que deseamos que en el texto se incluya.

5.Desarrollo de la herramienta

Para el estudio y análisis de la viabilidad de influir en el listado de Trending Topics o tendencias de una red social es necesario el desarrollo de una herramienta automática capaz de generar tráfico deseado con distintos significados semánticos y capaz de evitar ser detectada por spam. Es por esto último que nuestra herramienta debe permitir cierta flexibilidad semántica a la hora de publicar contenido en la red social. Además, al querer realizar los experimentos en distintas zonas geográficas la herramienta nos debe permitir cambiar el lugar desde el cual se desea publicar de una manera sencilla.

Por todo esto, los requisitos del sistema son:

- Librería capaz de conectar nuestra herramienta con la red social, en nuestro caso Twitter.
- Sistema capaz de generar lenguaje para evitar que nuestras cuentas sean etiquetadas como cuentas spammer y sean bloqueadas.
- Sistema capaz de ejecutar concurrentemente los distintos usuarios virtuales.
- Sistema que analice las redes en busca de si la herramienta ha cumplido su objetivo. Para ello nos basaremos tanto en la propia red social como en una página externa, trendsmap.com, que nos informa de las tendencias más importante en cada geolocalización en todo momento.

Una vez hemos visto tanto los requisitos que debe cumplir el sistema como la funcionalidad que debe cumplir la herramienta, vamos a ver el diagrama completo de esta, de tal forma que veamos los diferentes sistemas necesarios para cumplir con toda la funcionalidad de la que hablamos.

➢ ILUSTRACIÓN 9: SISTEMA COMPLETO

Como hemos comentado anteriormente, el objetivo del sistema es ser capaz de publicar en Twitter con distintos usuarios virtuales. Además, debemos de cumplir las normas que Twitter impone a sus usuarios en cuanto a spam, por lo que debemos ser capaces de generar distintas frases para cada uno de los usuarios cada vez que estos necesitan publicar. Realizando esta herramienta seremos capaces de enviar un gran volumen de mensajes a la red de una forma automática siendo capaces de ser una de las tendencias más populares del momento, cosa que sería imposible si se realizara manualmente.

En la imagen anterior se muestra las partes más básicas del proyecto, pudiéndose diferenciar en tres partes esenciales: el planificador, los usuarios y el comprobador.

- **Planificador:** Es la parte encargada del correcto funcionamiento del sistema. Su función se basa en el control de las cuentas de usuario de Twitter. Este sistema limita tanto el tiempo de vida de cada una de las cuentas virtuales como el tiempo entre publicaciones, de tal forma que no se incumplan las normas descritas por Twitter sobre uso indebido.

- **Sistema de usuarios:** Es la parte esencial del sistema. Es la parte encargada de la publicación en Twitter. Es en este sistema donde se utilizan los módulos de generación de lenguaje y de geolocalización. Se

encarga de las labores que debe realizar cada uno de los usuarios virtuales, publicar, retweetear...

- **Comprobador:** Este sistema nos proporciona una comprobación tanto en Twitter como en Trendsmap para ver si nuestro hashtag se encuentra entre los más populares del momento.

Además de los sistemas previamente comentados también contamos con una base de datos, como se muestra en la ilustración, que nos permite tanto guardar los datos relevantes de cada experimento realizado como la información necesaria para algunos sistemas.

Antes de ver cada uno de los sistemas en mayor detalle, tenemos que hablar del sistema de autenticación que utiliza Twitter para que las aplicaciones accedan al contenido de los usuarios.

5.1.Sistema de autenticación: OAuth

A partir del 31 de Agosto de 2010 las aplicaciones de terceros que se desarrollaron para la conexión a Twitter necesitaban el uso de OAuth para que las cuentas pudieran identificarse en la plataforma [23,24]. Este movimiento de autenticación en la plataforma mediante el sistema OAuth permitiría conexiones más fiables, seguras y cómodas para los usuarios.

OAuth es un protocolo de código abierto que permite la autorización segura de una API para aplicaciones de terceros, ya sea de escritorio, web o móviles mediante el uso del protocolo HTTP. Es decir, OAuth nos proporciona una comunicación sencilla entre la aplicación y el proveedor de servicios una vez el usuario a dado permiso a la primera para actuar en su nombre. Utilizando este protocolo lo que hacemos es permitir el acceso de la aplicación a nuestra cuenta de usuario mediante un clave de acceso (ó token), de tal forma que cada vez que intercambiamos información con el servidor no necesitamos usar nuestro usuario y contraseña, por lo que esta comunicación se vuelve más segura para el usuario.

En OAuth se definen cuatro roles distintos:

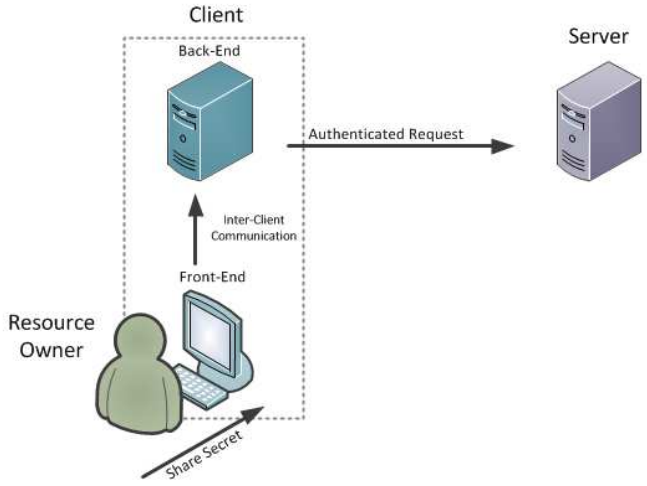
- **Propietario del recurso:** Es la entidad capaz de otorgar el acceso a los recursos protegidos. Cuando el propietario del recurso se refiere a una persona física a este se le conoce como usuario final (end-user).

- **Servidor de recursos:** El servidor que aloja los recursos protegidos. Este es capaz de aceptar y responder a las solicitudes de recursos protegidas utilizando la clave de acceso.

- **Cliente:** Se trata de una aplicación que hace peticiones de recursos protegidos al servidor de recursos en nombre de la propietario del recurso, con su previa autorización.

- **Servidor de autorización:** Es el servidor de emisión de claves de acceso al cliente. Posteriormente serán estas claves los que se utilizarán a la hora de solicitar recursos.

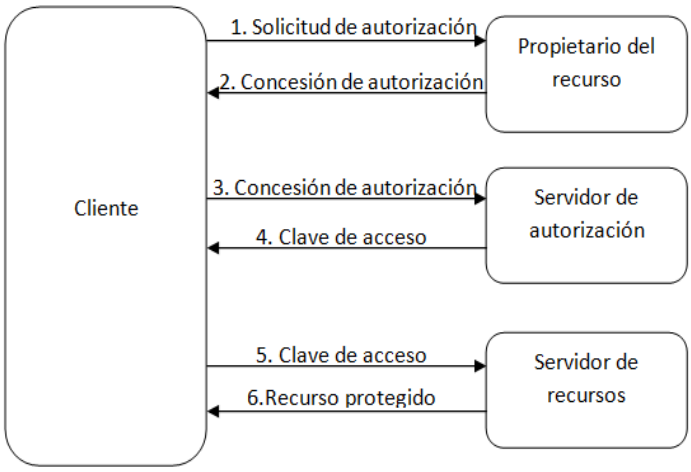
Generalmente, el servidor de autorización y el servidor de recursos se suelen representar como un único sistema por lo que el esquema de la arquitectura quedaría de la siguiente forma.



➢ ILUSTRACIÓN 10:SISTEMA OAUTH

En la figura anterior se muestra un esquema de la arquitectura de la tecnología. En ella podemos apreciar las tres partes de las que consta el sistema, el usuario, el administrador de la aplicación (en este caso simbolizado como una aplicación separada en su Back-End y Front-End) y por último el servidor de recursos y autorización.

El funcionamiento del sistema es el siguiente:



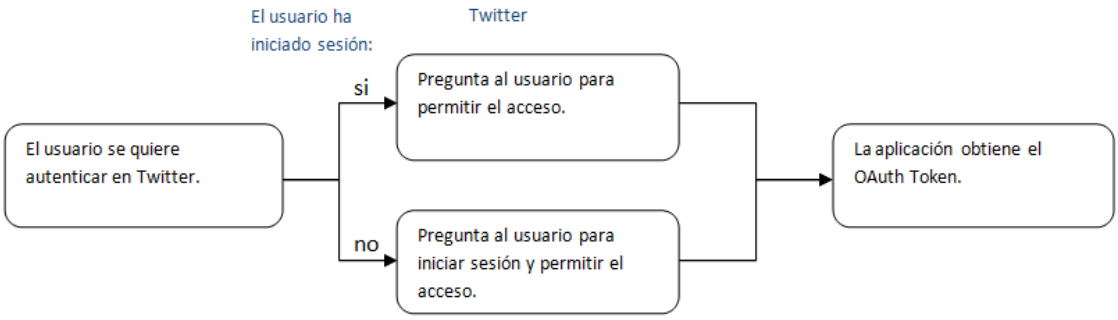
➢ ILUSTRACIÓN 11: PETICIÓN DE CLAVE, OAUTH

- El cliente solicita la autorización del propietario de los recursos.
- El cliente recibe una concesión de autorización, que es un credencial que representa la autorización del propietario del recurso.
- El cliente solicita una clave de acceso mediante la autenticación con el servidor de autorización y la presentación de la concesión de autorización.
- El servidor de autorización autentica al cliente y valida la concesión de autorización y, si es correcto, emite una clave de acceso para el cliente.
- El cliente solicita el recurso protegido del servidor de recursos mediante la presentación de la clave de acceso.
- El servidor de recursos valida el clave de acceso y, si es correcto, devuelve el contenido solicitado.

Como hemos visto, el encargado de realizar la autenticación es la aplicación, el usuario ya no necesita su usuario y contraseña para acceder al proveedor de servicios por lo que se vuelve más cómodo para el usuario la interacción.

El esquema previamente visto se aplica a las aplicaciones que acceden al contenido de Twitter. Es por ello que la aplicación es la que maneja todas las interacciones con el contenido de Twitter, tanto la publicación de nuevos tweets como la obtención de información de la página (ya sea la obtención del TimeLine, Trending Topics del momento o tweets asociados a un hashtag).

El procedimiento de autenticación de una cuenta mediante una aplicación es el siguiente:



➢ ILUSTRACIÓN 12:PETICIÓN DE CLAVE EN TWITTER, OAUTH

En la figura anterior se ve como se sigue el protocolo dictado por la RFC de OAuth. En primer lugar se manda una petición al servidor con la clave de acceso de la aplicación creada. Una vez ha sido recibida se solicita la información del usuario, que si ya ha iniciado sesión no será necesaria, y se le pide la confirmación para que la aplicación actúe en su nombre. Cuando el usuario ya ha dado su confirmación, el servidor mandará al usuario una clave para dar de alta a la aplicación en su cuenta que la aplicación deberá enviar al servidor. Una vez este procedimiento ha sido realizado, el servidor enviará una clave de acceso para acceder a los recursos protegidos. Una vez la aplicación tiene esta clave, esta será de uso ilimitado, por lo que este procedimiento solo será necesario realizarlo una única vez.

5.2. Planificador

Como hemos comentado previamente es la parte esencial del programa, la cual se encarga del manejo de todos los usuarios. Esta se puede dividir en dos partes a su vez: el generador de distribución y el planificador propiamente dicho.

El generador de distribución es la función encargada de definir los tiempos de inicio y fin como de los intervalos de publicación de mensajes mediante los cuales se simulará una distribución en la publicación de tweets. Esta función se ejecuta al inicio del programa, incluyendo en una tabla los tiempos previamente descritos, que servirán a los posteriores sistemas saber cómo deben actuar.

El planificador es el encargado del manejo de los usuarios propiamente dicho. Con los tiempos de actuación de cada uno de los usuarios que han sido definidos en el generador de distribución, el planificador se encarga de eliminar los usuarios y de crearlos en los tiempos estipulados.

5.2.1.Generador de distribución

Para generar las distribuciones se ha optado por generar dos tipos distintos que nos permitan generar distintos patrones de conducta: uno de carácter lineal y otro de carácter exponencial. El primero se ha elegido debido a la simplicidad de implementación y el segundo debido a que la mayoría de los *Trending Topic* siguen este tipo de distribución. Pese a todo, ya que ambos tipos de distribución pueden resultar en un *Trending Topic*, el generador será capaz de generar ambos tipos de distribución.

Para ambos tipos de distribución se han tenido en cuenta dos aspectos importantes para generar correctamente la distribución:

- **Aleatorización del inicio:** Con esto se refiere a que cada una de las cuentas sujeta a un mismo patrón de activación no inicia su ejecución en el mismo momento. Como se verá más adelante, se definirán varios grupos de activación, en el que un determinado número de usuarios estarán activos durante un periodo determinado de tiempo. Pese a que todos ellos actúan de forma igual una vez han iniciado su ejecución, el tiempo de activación ha sido aleatorizado de tal forma que no todas las cuentas comiencen su ejecución en el mismo momento. Este proceso se realiza para evitar que Twitter identifique las cuentas que van a empezar a publicar como spammer, ya que todas ellas estarán publicando con el mismo hashtag durante un periodo de tiempo prolongado.

- **Tiempos de espera:** dentro de cada una de los grupos de usuarios veremos dos números asociados al tiempo de espera. Estos tiempos son los tiempos entre los cuales el usuario permanecerá sin publicar en la red. Se definen dos números: el tiempo de espera mínimo y el tiempo de espera máximo. Como se ha comentado anteriormente, se aleatorizó el tiempo de inicio con el fin de evitar ser detectado como spammer, en este caso haremos lo mismo con el tiempo de publicación, de tal forma que pese a que las cuentas en un determinado momento publiquen en un mismo momento, debido al tiempo de espera aleatorio es poco probable que vuelvan a coincidir. Además, incluyendo este componente aleatorio evitamos que se nos detecte debido a la publicación arbitraria de tweets cada cierto periodo de tiempo.

Una vez aclarados dos de los aspectos fundamentales para el generador de distribución, vamos a ver las peculiaridades de ambos tipos de distribuciones.

Distribución de carácter lineal

Como hemos comentado anteriormente, esta distribución es la más sencilla de las dos propuestas. Esta se basa en que todas las cuentas estarán activas durante el periodo de tiempo designado. Además, está diseñado para que todas las cuentas tengan el mismo peso en la generación de la distribución, es decir, todas y cada una de las cuentas tendrán unos tiempos de espera iguales entre publicaciones.

A continuación se muestra una tabla con la información de activación de cuentas junto a una simulación de la distribución generada.

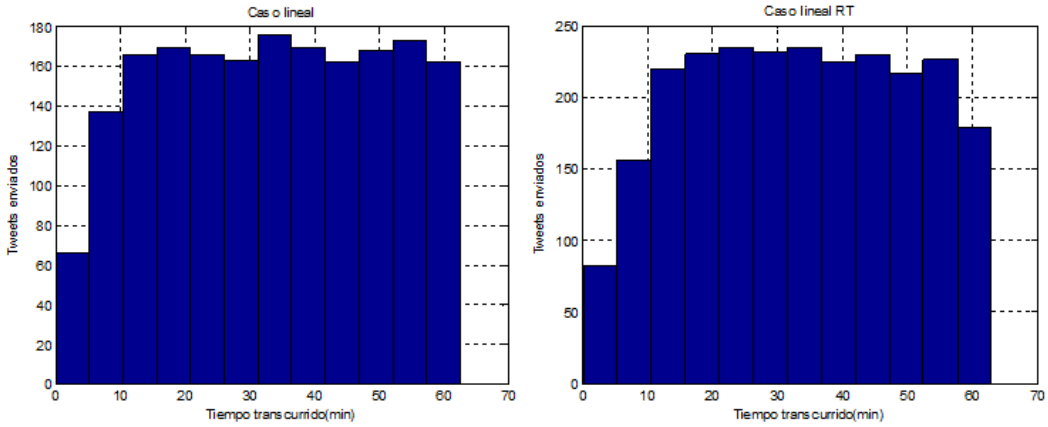
➢ TABLA 1: INFORMACIÓN DISTRIBUCIÓN LINEAL

Grupo	# Usuarios	T. Inicio	T. Espera	T. Fin
Grupo A	80	0	2-3	60

Como vemos en la tabla, todas las cuentas de usuario pertenecen al mismo grupo, como dijimos previamente, y estas se ejecutan durante la totalidad del experimento.

A continuación se mostrará la simulación realizada con los datos proporcionados en la tabla anterior.

Lineal



➢ ILUSTRACIÓN 13: SIMULACIÓN DISTRIBUCIÓN LINEAL

En la ilustración podemos apreciar dos matices importantes:

- El primer punto a tener en cuenta es el inicio. Como bien hemos dicho antes, se ha aleatorizado el inicio de todas las cuentas, de esta forma vemos en la gráfica como, al inicio de la emisión de tweets el número es menor que cuando se ha estabilizado.

- El segundo punto son las pequeñas variaciones en el número de tweets. Esto es debido al carácter aleatorio de la emisión en los tiempos de espera, como bien se comentó anteriormente.

Distribución de carácter exponencial

Es el segundo tipo de distribuciones que podremos generar con nuestra sistema. De este tipo de distribuciones generaremos dos clases distintas: la propiamente exponencial y la de carácter intensivo.

El carácter exponencial de estas distribuciones es un poco más complicado de generar mediante el conjunto de usuarios, por lo que se procederá a crear varios grupos, los cuales seguirán unas mismas normas de publicación. De esta forma al ver los resultados de la simulación podremos ver qué aspectos de la distribución no se ajustan a la realidad, pudiendo cambiar un único grupo y no todo el conjunto de usuarios. Para ver el desarrollo de la distribución vamos a ir viendo la progresión de los diferentes grupos y su comportamiento en la distribución.

Aquí solo veremos los pasos más importantes de la evolución que han sufrido estas distribuciones, para que entendamos mejor el comportamiento que deseamos en estas.

Versión uno: datos arbitrarios.

En este primera versión partimos sin previo ejemplo, por lo que el número de grupos como el tiempo de ejecución de estos son arbitrarios, intentando que el grueso del número de publicaciones se produzca al final de la ejecución. Sin más, vamos a ver los datos de esta primera aproximación.

A continuación mostramos los grupos virtuales de usuarios que se han creado y sus principales características.

➢ TABLA 2: INFORMACIÓN DISTRIBUCIÓN EXPONENCIAL 1

Grupo	# Usuarios	T. Inicio	T. Espera	T. Fin
A	$\frac{Nusuarios}{10}$	0	$2-6$	$totalTime$
B	$\frac{Nusuarios \cdot 2}{10}$	$\frac{totalTime}{3}$	$3-5$	$totalTime$
C	$\frac{Nusuarios \cdot 3}{10}$	$\frac{totalTime}{2}$	$2-4$	$totalTime$
D	$\frac{Nusuarios \cdot 4}{10}$	$\frac{totalTime}{2}$	$1-3$	$\frac{3 \cdot totalTime}{4}$

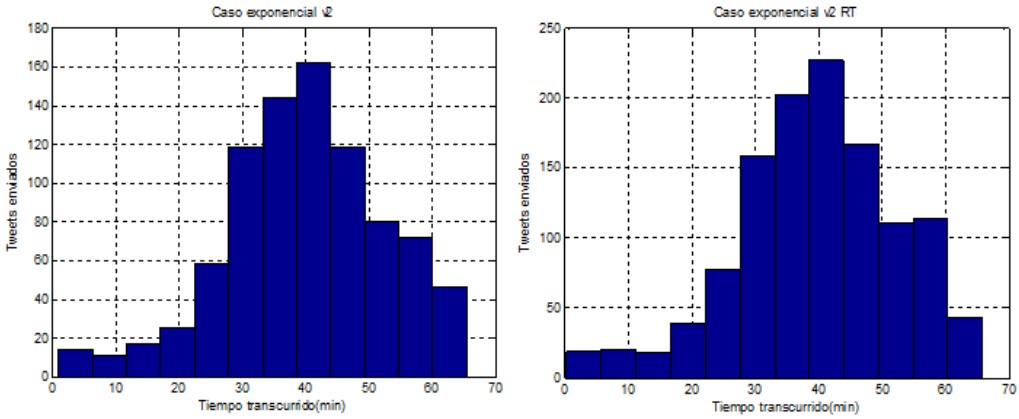
Como vemos en la tabla anterior, el grueso de los usuarios se encuentran en los grupos C y D, donde a su vez son los que tienen un periodo más corto de espera entre publicaciones. El fin de realizar esto es intentar que el número máximo de mensajes publicados en Twitter se realice al final de la ejecución.

Otra de las peculiaridades que podemos apreciar es que todos los grupos excepto el D tienen su finalización con el tiempo indicado por parámetro. Haciendo que el grupo D acabe un poco antes su ejecución conseguimos que al final de nuestra distribución tenga una ligera caída de publicaciones, que intenta simular la pérdida de interés de los usuarios en el hashtag utilizado.

Por último, decir que *Nusuarios* hace referencia al número de usuarios con los que se desea publicar en Twitter. De esta forma, podemos realizar distintos experimentos con distintos usuarios sin modificar el código. Este modo de distribución de usuarios será también aplicado en las siguientes evoluciones de esta distribución.

Sin más, vamos a ver las simulaciones realizadas con este tipo de configuración de usuarios.

Versión 1



➢ ILUSTRACIÓN 14:SIMULACIÓN DISTRIBUCIÓN EXPONENCIAL 1

En las gráficas tenemos dos simulaciones, una que solo incluye los mensajes publicados por las cuentas, la de la izquierda, y la otra que muestra también los posibles retweets que se realizarán con estas cuentas. Cada una de las barras indica el número de publicaciones que se realizarían cada 5 minutos de ejecución. Esta simulación esta realizada con un número de usuarios total de 80 usuarios, que es el máximo número del que se dispone.

A continuación vamos a analizar detenidamente las simulaciones generadas y ver qué puntos son necesarios modificar para que la distribución se parezca más a una que pudieran generar los humanos.

- En primer lugar, remarcar que el inicio de la simulación es muy lenta, es decir, el crecimiento del número de mensajes es algo lento al inicio de esta.

- El segundo punto a recalcar es el abrupto crecimiento de las publicaciones durante la fase intermedia, entre los minutos 25 y 35.

- Por último, el rápido desplome de las publicaciones una vez el grupo D de usuarios termina su ejecución.

A pesar de los problemas previamente comentados, vemos como la distribución generada es una distribución exponencial por lo que partimos de una buena base.

Versión dos: crecimiento mejorado

Una vez hemos visto el primer paso de la generación de la distribución vamos a realizar algunos cambios en el reparto de usuarios en grupos.

El primer cambio que vamos a realizar es la reutilización de usuarios, es decir, los usuarios podrán ser asignados a dos grupos distintos, aunque estos no comparten tiempos de ejecución entre sí. De esta forma, estaremos cambiando el régimen de publicación de cada una de las cuentas involucradas en el experimento.

El segundo cambio que se realizará es la creación de un nuevo grupo, el grupo E, que compartirá tiempo de ejecución con el grupo C, pero que su tasa de envío de mensajes es algo más lenta.

Por último se han modificado algo tanto los tiempos de espera como los de inicio y fin de algunos de los grupos, para intentar que el crecimiento de nuestra distribución no sea tan abrupto como lo era en el caso anterior.

Así la tabla de usuarios de esta nueva versión sería la siguiente:

➢ TABLA 3:INFORMACIÓN DISTRIBUCIÓN EXPONENCIAL 2

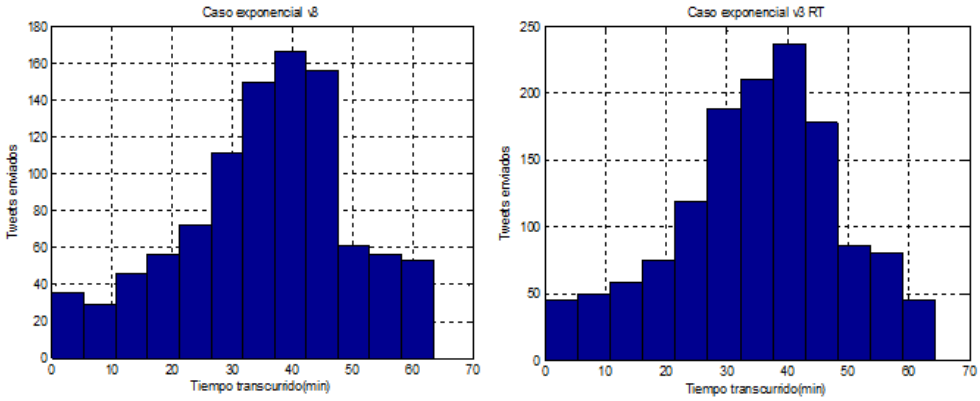
Grupo	# Usuarios	T. Inicio	T. Espera	T. Fin
A	$\frac{Nusuarios \cdot 3}{10}$	0	$3 - 6$	$\frac{totalTime}{2}$
B	$\frac{Nusuarios \cdot 2}{10}$	$\frac{totalTime}{4}$	$2 - 4.5$	$\frac{3 \cdot totalTime}{4}$
C	$\frac{Nusuarios \cdot 2}{10}$	$\frac{totalTime}{2}$	$2 - 4$	$totalTime$
D	$\frac{Nusuarios \cdot 4}{10}$	$\frac{totalTime}{2}$	$1 - 3$	$\frac{3 \cdot totalTime}{4}$
E	$\frac{Nusuarios \cdot 2}{10}$	$\frac{totalTime}{2}$	$3 - 4$	$totalTime$

Como hemos dicho previamente, al sumar el total del número de usuarios este es superior al indicado por el experimento, pero las cuentas involucradas en el mismo serán reutilizadas en los diferentes grupos mencionados.

Podemos apreciar a su vez como el tiempo de espera del grupo A ha aumentado y el del grupo B a disminuido ligeramente. Con esto dos cambios se intenta que el crecimiento de nuestra distribución sea algo más suave que en la versión previa.

Una vez analizadas las tablas de usuarios, vamos a ver las simulaciones correspondientes a esta versión de la distribución.

Versión 2



➢ ILUSTRACIÓN 15: SIMULACIÓN DISTRIBUCIÓN EXPONENCIAL 2

En las gráficas tenemos dos simulaciones, una que solo incluye los mensajes publicados por las cuentas, la de la izquierda, y la otra que muestra también los posibles retweets que se realizarán con estas cuentas. Cada una de las barras indica el número de publicaciones que se realizarían cada 5 minutos de ejecución. Esta simulación esta realizada con un número de usuarios total de 80, que es el máximo número del que se dispone.

Analizando los datos provistos por la gráfica vemos que el problema del abrupto crecimiento que la distribución sufría ha sido solucionado aunque el rápido decrecimiento del volumen de mensajes una vez alcanzado el máximo sigue siendo un problema. Para solucionar esto vamos a ver la tercera versión de la distribución y definitiva.

Versión tres: decrecimiento mejorado

Como anteriormente se ha explicado, aun nos queda por solucionar el problema de el rápido decrecimiento que sufre nuestra distribución una vez ha alcanzado su máximo. Nuestro objetivo es que el decrecimiento del volumen de datos generado por nuestras cuentas sea algo más suave y para ello aplicaremos la misma técnica que hemos comentado previamente, añadir un nuevo grupo de usuarios, en este caso el F. Este nuevo grupo nos permitirá hacer una transición más suave para que el decrecimiento sea más suave.

Sin más, vamos a ver la tabla de distintos grupos propuesta para esta nueva versión de la distribución.

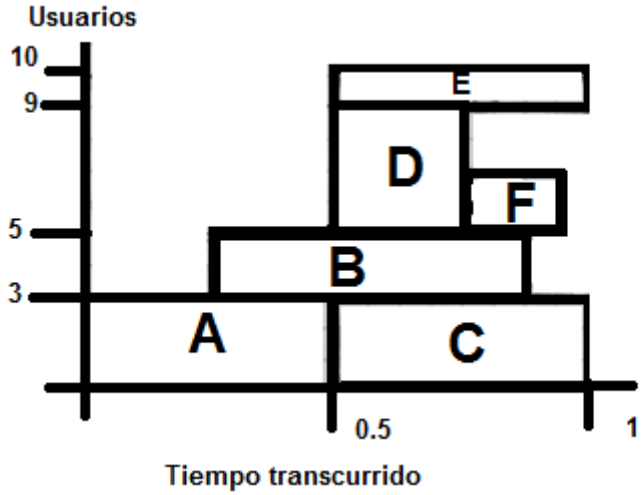
➢ TABLA 4:INFORMACIÓN DISTRIBUCIÓN EXPONENCIAL 3

Grupo	# Usuarios	T. Inicio	T. Espera	T. Fin
A	$\frac{Nusuarios \cdot 3}{10}$	0	$4-8$	$\frac{totalTime}{2}$
B	$\frac{Nusuarios \cdot 2}{10}$	$\frac{totalTime}{4}$	$2-4$	$\frac{7 \cdot totalTime}{8}$
C	$\frac{Nusuarios \cdot 3}{10}$	$\frac{totalTime}{2} + A$	$2-4$	$totalTime$
D	$\frac{Nusuarios \cdot 4}{10}$	$\frac{totalTime}{2}$	$1-3$	$\frac{3 \cdot totalTime}{4}$
E	$\frac{Nusuarios}{10}$	$\frac{totalTime}{2}$	$3-4$	$totalTime$
F	$\frac{Nusuarios \cdot 2}{10}$	$\frac{3 \cdot totalTime}{4} + D$	$4-8$	$\frac{15 \cdot totalTime}{16}$

Como vemos en la tabla, el número de usuarios en cada grupo varia ligeramente, para que estos se acoplen mejor a la nueva distribución.

Otro punto a tener en cuenta es el tiempo de actuación que tiene el nuevo grupo F. Como vemos este tiempo es muy reducido, entre $3/4$ y $15/16$ del tiempo total de la simulación. Como se dijo previamente, el objetivo de este grupo de usuarios es que la distribución no sea tan abrupta al final de su ejecución, por lo que no necesitamos que este grupo actúe en mas etapas de la simulación.

Por último, me gustaría destacar los tiempos de inicio de los grupos F y C, donde se incluyen los tiempos A y D respectivamente. Como se dijo en la versión anterior de la distribución, las cuentas de usuarios son utilizadas en distintos grupos, con el fin de que publiquen con otro régimen distinto. Para realizar esto correctamente debemos de cuidar los tiempos de iniciación de algunos de nuestros grupos para que estos no se solapen y pueda llegar a ocurrir un error.

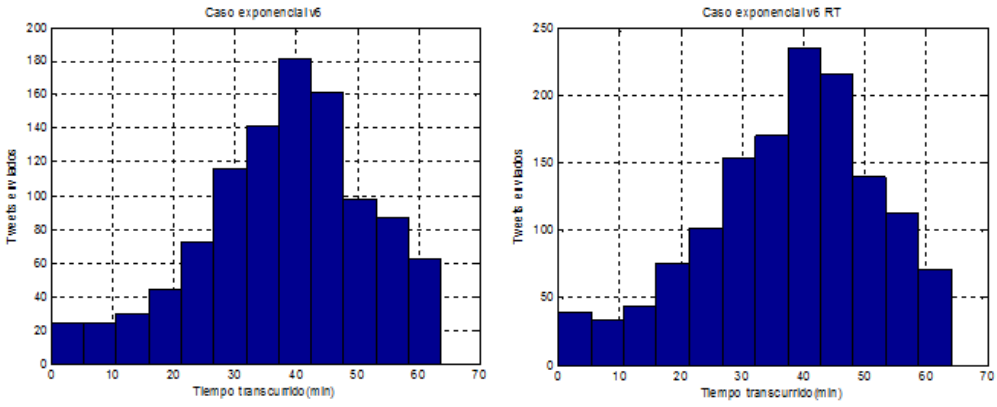


➢ ILUSTRACIÓN 16:DIAGRAMA DE USUARIOS

En la imagen anterior se muestran los tiempos de ejecución de las cuentas, normalizadas a diez usuarios. Como vemos, los tiempos de ejecución de estos se complementan, de tal forma que los grupos A y C y D y F utilizarán los mismos usuarios pero con regímenes de publicación totalmente distintos, de tal forma que la distribución generada se amolde a las características que deseamos de ella.

Una vez más, vamos a ver las simulaciones generadas por esta versión de la distribución.

Versión 3



➢ ILUSTRACIÓN 17: SIMULACIÓN DISTRIBUCIÓN EXPONENCIAL 3

En las gráficas tenemos dos simulaciones, una que solo incluye los mensajes publicados por las cuentas, la de la izquierda, y la otra que muestra también los

posibles retweets que se realizarán con estas cuentas. Cada una de las barras indica el número de publicaciones que se realizarían cada 5 minutos de ejecución. Esta simulación esta realizada con un número de usuarios total de 80, que es el máximo número del que se dispone.

Como vemos, el problema que intentábamos solventar ha sido solucionado, ahora el decrecimiento del volumen de datos generados ahora es menos abrupto, simulando mejor el comportamiento que puede tener un hashtag creado por los seres humanos.

Distribución de carácter exponencial: intensiva

Una vez hemos visto la distribución de carácter exponencial y sus características, ahora vamos a ver otro tipo de distribución, también exponencial, pero con otro tipo de características.

Esta distribución es llamada intensiva. Está diseñada para experimentos de larga duración (a partir de 12 horas) y es la que utilizaremos para intentar conseguir que nuestro hashtag aparezca como Trending Topic.

El funcionamiento de esta distribución es muy sencillo y tiene las siguientes características:

- Lo primero de todo, esta distribución es de carácter exponencial, es decir, el crecimiento del flujo de mensajes cumple con este tipo de distribución.

- Las cuentas de usuarios, a diferencia de la distribución anterior, se van añadiendo una a una, y no mediante bloques.

- Una vez el número de usuarios máximo se ha alcanzado estos se mantendrán enviando mensajes hasta la finalización del experimento. Con esto se quiere decir que, una vez se ha alcanzado el número máximo de usuarios, el número de mensajes que se publicarán a lo largo del tiempo será constante.

Una vez hemos visto las características más importantes de esta distribución, vamos a ver cómo funciona.

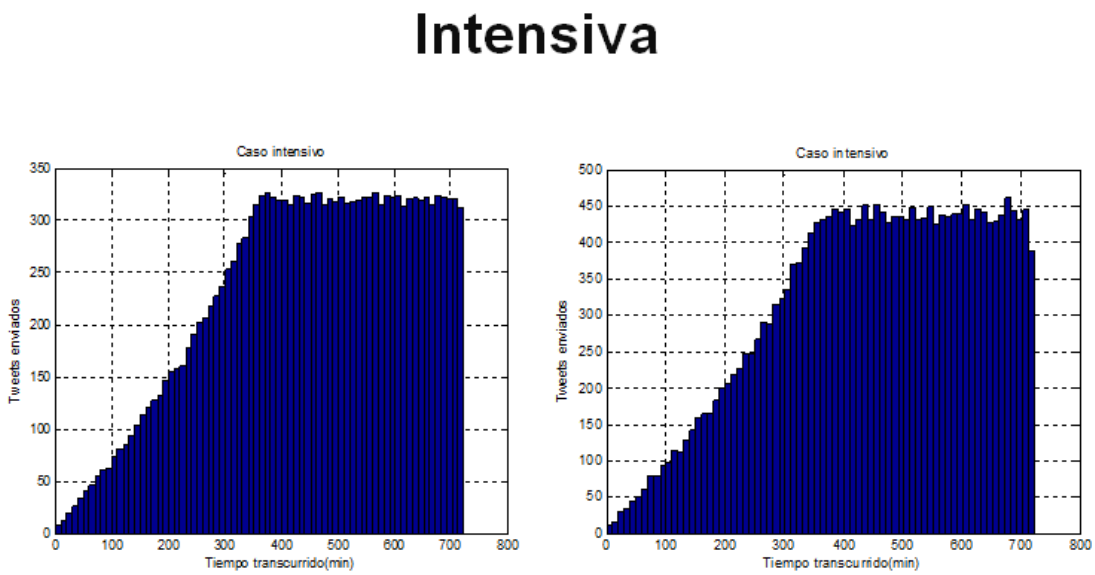
- En primer lugar, todas las cuentas de usuarios publicarán con los mismos tiempos mínimos y máximos, de esta forma es más sencillo regular el tipo de distribución que se está generando, además de que es más sencillo de

controlar que las cuentas no sobrepasen el límite de publicaciones impuesto por Twitter.

- El tiempo entre creación de usuarios depende del número de usuarios previamente generados. De esta forma conseguimos que, a medida que va pasando el tiempo, el volumen de datos generado sea mayor cuanto más tiempo pasa. El tiempo entre generación de usuarios es de 6 minutos al inicio y los últimos se generan con un frecuencia de 2 minutos únicamente.

- Por último, el tiempo final de la ejecución de las cuentas es el mismo para todas ellas y coincide con el tiempo indicado como tiempo máximo de ejecución.

Una vez hemos visto como funciona, vamos a ver dos simulaciones de la distribución. En la primera de ellas se verán los mensajes creados por los usuarios mientras que en la segunda figura realizaremos una simulación añadiendo un factor de retweet, que es lo que se utilizará en el experimento final.



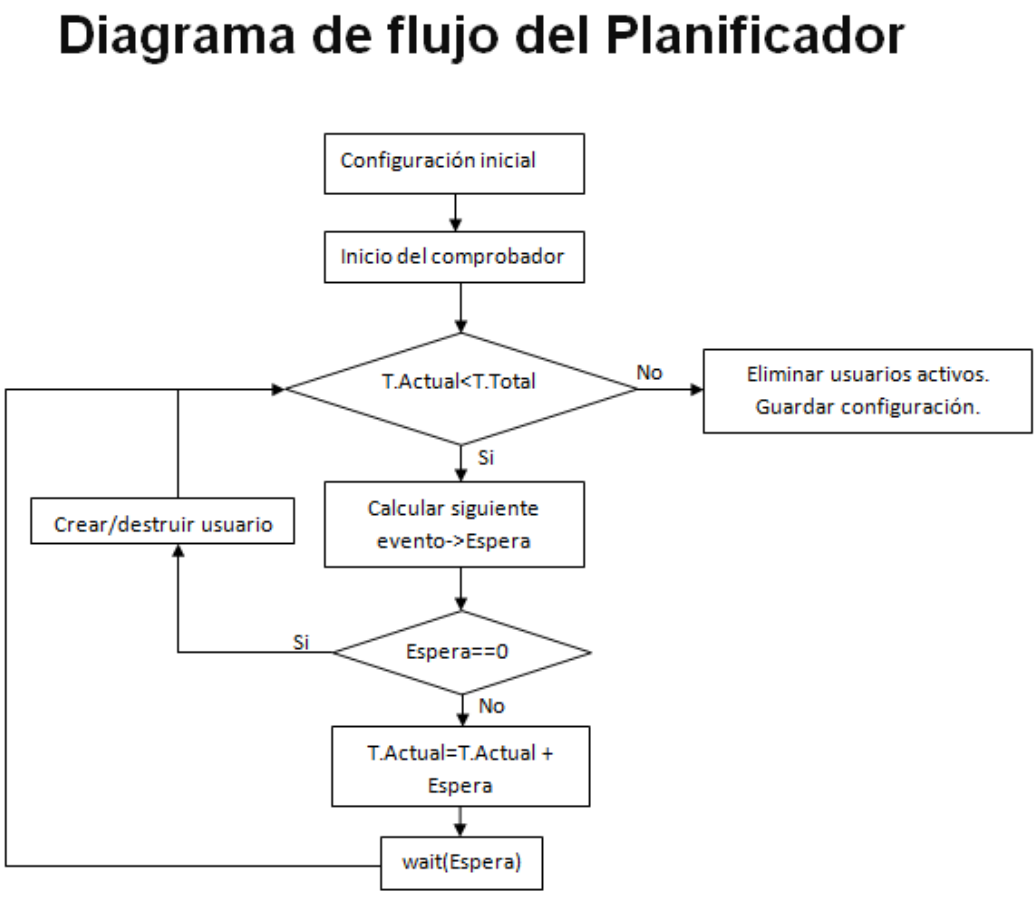
➢ ILUSTRACIÓN 18: SIMULACIÓN DISTRIBUCIÓN INTENSIVA

A la izquierda tenemos la simulación sin el coeficiente de retweet y a la derecha con este. Como vemos en las imágenes el crecimiento de ambas simulaciones es mayor a medida que va pasando el tiempo y una vez este máximo de cuentas se ha alcanzado este máximo se queda estable.

Podemos también apreciar como en el caso de tener retweets, el crecimiento es algo más irregular y una vez se ha alcanzado el máximo de usuarios, el volumen de mensajes es algo más irregular que en el caso de no tenerlo.

5.2.2. Planificador

Una vez hemos visto el comportamiento del generador de distribuciones así como los diferentes tipos que se pueden generar, vamos a ver el planificador, que es la parte encargada de que cada uno de los usuarios cumpla con su periodo de vida correspondiente. En primer lugar vamos a ver el diagrama de flujo de este sistema y posteriormente lo analizaremos paso a paso.



➢ ILUSTRACIÓN 19:DIAGRAMA DE FLUJO DEL PLANIFICADOR

Vamos a analizar detenidamente cada uno de los pasos que seguirá nuestro planificador:

- En primer lugar, se crea el planificador pasándole todos los datos necesarios para su correcto funcionamiento. Entre los datos necesarios tenemos: nombre del experimento, distribución, número máximo de usuarios, tiempo total de ejecución, hashtag utilizado y localización. Toda esta información será utilizada durante su ejecución tanto para la

administración de tiempos de vida de las cuentas de usuario así como para la correcta inicialización de los usuarios.

- El siguiente paso es la inicialización del propio planificador. En este paso se realiza la comprobación de la localización, la creación de la conexión con la base de datos, que nos permitirá enviársela a los usuarios creados, la creación del comprobador y la inicialización del generador de lenguaje.

- Una vez realizada la configuración del planificador, entramos en la ejecución propia del planificador. Para ello, hemos creado un bucle donde se compara el tiempo relativo del planificador con el tiempo total de la ejecución. En caso de que el tiempo relativo sea menor que el tiempo total, el planificador seguirá su ejecución. En caso contrario esto significa que el tiempo total del experimento ha acabado, y se procederá a la finalización de todas las cuentas que se encuentren aún activas.

- Estando en el bucle, la primera tarea que realizará el planificador es comprobar la siguiente tarea que debe realizar. Para ello, comprobará la lista de tiempos de inicio y fin de cada una de las cuentas que se necesitan en el experimento. En caso de que el tiempo de espera sea 0, es decir, se deba realizar la tarea en ese mismo instante, se buscará si la tarea a realizar es creación o destrucción de un usuario. En ambos casos, hay que tener cuidado con la actualización de los datos guardados en la base de datos, ya que debemos hacer que estas cuentas de usuario sean bloqueadas o liberadas, en función de si es creación o destrucción respectivamente.

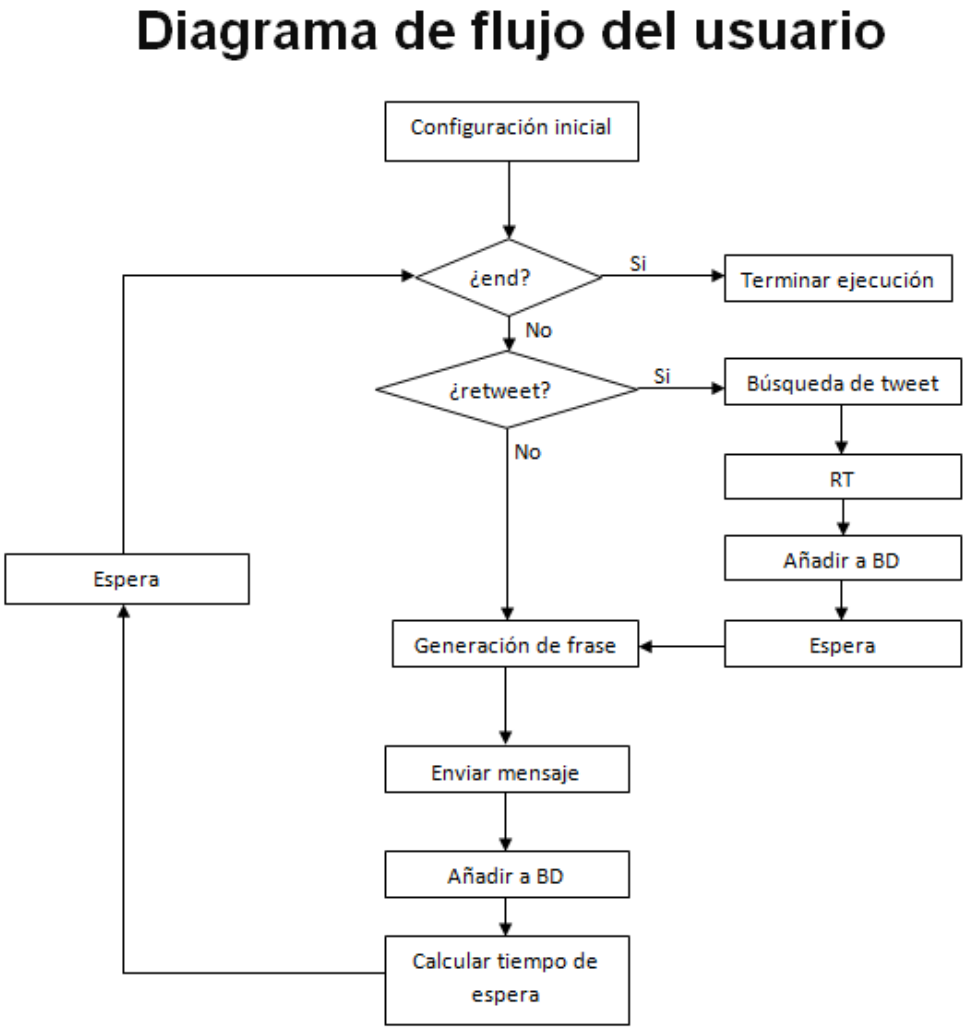
- En el caso de que la siguiente tarea a realizar sea en un futuro, es decir, que el tiempo de espera sea distinto de 0, el hilo encargado de planificar las cuentas realizará una espera determinada por el tiempo hasta la siguiente tarea.

Como hemos visto en el diagrama de flujo del planificador, su ejecución es bastante sencilla, ya que la única tarea que debe supervisar es el tiempo de nacimiento y muerte de cada uno de los hilos encargados de los usuarios.

5.3. Sistema de usuarios

Es la parte esencial de nuestro sistema completo. Este sistema se encarga de el control de cada una de las cuentas de usuario de forma individualizada para que estas publiquen en Twitter dependiendo de los datos proporcionados por el planificador, tiempo entre publicaciones y tiempo de inicio y final de ejecución.

Para que nos hagamos una idea del funcionamiento general de este sistema vamos a ver el diagrama de flujo que sigue cada una de las cuentas de usuario para publicar en Twitter.



➢ ILUSTRACIÓN 20:DIAGRAMA DE FLUJO DEL USUARIO

A continuación, vamos a explicar el proceso que cada uno de los usuarios seguirá durante su tiempo de vida.

- Lo primero que se hará es la configuración de cada uno de los usuarios, obteniendo información sobre geolocalización, base de datos, tiempos de

espera máximos y mínimos, datos de acceso a Twitter y acceso a el sistema de lenguaje, esto está representado como la configuración inicial.

- Una vez realizado este proceso de inicialización entraremos en un bucle determinado por un flag (en este caso end) que será el flag que permitirá a nuestro planificador decir a cada uno de nuestros usuarios cuando debe terminar su ejecución. Es por ello que, hasta que el planificador no diga lo contrario nuestro usuario seguirá su ejecución.

- Mientras el planificador permita la ejecución del usuario, el usuario podrá tanto realizar retweets así como publicar sus propios tweets. En nuestro caso, el orden de ejecución será que primero se comprobará si se decide o no retweetear y tras ello se publicará un tweet propio.

- Para la función de retweet se comparará un parámetro a un número generado aleatoriamente para decidir si el usuario, en esta iteración, retweeteará algún mensaje previamente enviado. En caso de que se realice el retweet, el usuario deberá esperar un pequeño tiempo, que servirá para no superar el límite de publicación impuesto por Twitter. Esta espera estará entre los 24 y 36 segundos.

- Tanto en el caso de que la cuenta haga retweet o no, la cuenta publicará un mensaje propio. Para ello el usuario llamará al generador de lenguaje para que este le proporcione el mensaje que se enviará.

- Una vez se ha obtenido el mensaje que se va a enviar (este ya incluirá el hashtag escogido al inicio de la ejecución) este se publicará mediante el API de Twitter4j.

- Por último, se realizará la espera necesaria que nos indica el planificador con sus tiempos máximos y mínimos de publicación y se volverá al inicio del bucle comentado anteriormente.

Además del funcionamiento general de cada uno de los usuarios, cabe destacar tres puntos adicionales en la ejecución de este código:

- El primer concepto es la base de datos de tweets enviados. Esta pequeña base de datos se guardará en memoria RAM, por lo que esta será independiente en cada una de las ejecuciones y será eliminada tras acabar el experimento. Esta se guarda en RAM debido a que en cada una de las ejecuciones del sistema el hashtag que se va a usar es distinto, por lo que el almacenamiento de estos tweets para posterior uso es inútil.

La principal función de esta pequeña base de datos es guardar los tweets asociados a la ejecución para que cada uno de los usuarios pueda realizar retweets de mensajes previamente enviados con el hashtag deseado.

La otra opción para realizar la operación de reenvío de mensajes seria la búsqueda mediante el API de Twitter4j en la base de datos de Twitter del hashtag que estamos utilizando en ese momento. Por desgracia, debido a la limitación impuesta de Twitter para las aplicaciones, si el número de cuentas con las que cuenta una aplicación es muy elevada, Twitter no responderá a todas las peticiones que se le manden. Es por esto que, para permitir un margen de retweet variable que no impida el margen provisto por Twitter se debe utilizar esta base de datos.

- El segundo concepto a tener en cuenta en el código de los usuarios es la selección del mensaje que se desea hacer retweet. Para la selección de los mensajes se tienen en cuenta dos factores, el usuario y el tiempo cuando se envió.

  En cuanto al usuario, debemos comprobar que el mensaje que se intenta hacer retweet no ha sido previamente enviado por nuestra cuenta ya que en caso de que este mensaje se haya mandado previamente Twitter nos enviará un mensaje de error. Es por ello que se deberá comprobar si los usuarios han enviado ya dicho mensaje.

- El otro aspecto importante, el tiempo de envío, es un factor importante. Para seleccionar el mensaje que se desee enviar lo seleccionaremos de tal forma que los mensajes recientes tengan una mayor probabilidad de ser seleccionados. Esto se hace para emular un poco más el comportamiento humano. Generalmente los mensajes mas retweeteados son los que están más alto del TimeLine de cada uno de los usuarios, ya que son los más accesibles para ellos. La ecuación para seleccionar el tweet es el siguiente:

  $$index = (1 - Math.random() \cdot Math.random()) \cdot BD.size$$

  Es decir, generamos dos números aleatorios que se multiplican entre sí, y el producto de estos es restado a uno. Después, en función del tamaño de la base de datos de tweets que tengamos se elegirá uno en la posición correspondiente. Con esta ecuación se intenta que los tweets enviados más recientemente tengan una mayor probabilidad de que se seleccione dicho mensaje para ser retweeteado.

  Por último, cabe destacar que cada vez que se hace un retweet a un usuario, el usuario que reenvía el mensaje empieza a seguir al usuario del tweet original. De esta forma, todos los usuarios de nuestro sistema,

poco a poco, se van a ir siguiendo entre ellos, de tal forma que la repercusión total de las cuentas en la red irá aumentando a medida que las cuentas se vayan haciendo retweet.

Para que cada uno de los usuarios cumpla su objetivo vamos a utilizar dos librerías externas que nos proporcionarán la funcionalidad básica para que nuestro sistema sea eficiente. Estas dos librerías son: Twitter4j y SimpleNLG.

Además de estas dos librerías necesitaremos utilizar la geolocalización para vincular nuestro mensaje a una zona geográfica determinada.

5.3.1. Geolocalización

La geolocalización se trata de un sistema que permite saber las coordenadas geográficas de un objeto en cualquier posición de la Tierra [25]. Para conocer la posición del sistema a localizar se utiliza un sistema de radio frecuencia, generalmente TDOA (Time Diference Of Arrival) que mediante el tiempo de trayecto de varias señales desde distintos puntos (generalmente satélites de GPS) miden la distancia entre el objeto y los puntos, con lo que el objeto se puede triangular mediante estas medidas.

Con el auge de los teléfonos móviles y tabletas, las redes sociales decidieron añadir la función de geolocalización, lo que permite añadir la localización exacta en la que te encuentras al mensaje que envías. Algunas de las más famosas redes sociales que han implementado esta funcionalidad son Facebook o Twitter.

Como se comento previamente, Twitter tiene un sistema llamado Trending Topics que permite saber cuáles son las tendencias en cada momento. Con la ayuda del sistema de geolocalización, estos Trending Topics pueden ser mas locales, es decir, puedes saber cuáles son los temas más hablados en tu zona y no solo en el mundo. Este es un método muy interesante sobre todo para conocer noticias regionales, ya que aunque no tengan la suficiente repercusión como para llegar a ser un Trending Topic mundial sí que lo pueden ser en tu zona.

Para añadir la información de geolocalización en Twitter existen dos formas, una mediante la dirección IP desde la cual se manda el Tweet, que mediante una petición WHOIS permite saber la zona en la que dicha IP esta asignada, o mediante la incorporación de las coordenadas geográficas dentro del mensaje.

Generalmente, al ser en teléfonos móviles donde más se utiliza la función de geolocalización, es el segundo caso el más utilizado.

Como se ha comentado previamente, Twitter ofrece un sistema de Trending Topic local, que permite saber cuáles son las tendencias en una zona determinada. Para ello, Twitter utiliza el sistema de Yahoo WOEID (Where On Earth IDentifier), que ofrece un código identificativo a cada zona de la Tierra. De tal forma que, los mensajes enviados en cada una de las zonas geográficas de un código WOEID se analizaran para saber cuáles son las tendencias de dicha zona.

5.3.2. Twitter4j

Gracias al API-rest que proporciona el propio Twitter existe una gran cantidad de desarrolladores que han incluido funcionalidad dentro de sus aplicaciones.

Basados en el API-rest que proporciona Twitter, terceras personas han desarrollado librerías propias para simplificar la comunicación con Twitter. Entre ellas esta Twitter4j, que es la librería que utilizaremos entre proyecto para realizar la comunicación de Twitter.

Twitter4j es un librería escrita en java que permite la comunicación con el servidor de Twitter [9]. Esta librería nos proporciona un gran abanico de utilidades:

- Librería totalmente basada en Java.
- Librería totalmente independiente de ficheros externos.
- Soporte para la comunicación mediante el protocolo OAuth necesario para la comunicación entre aplicaciones y los servidores de Twitter.
- Totalmente compatible con última versión del API de Twitter.

Utilizando Twitter4j se simplifica en gran medida la comunicación entre nuestro sistema y Twitter, ya que con el uso de las clases definidas dentro de la librería nos permite modificar cualquiera de las propiedades, ya sean de los tweets enviados o de la misma cuenta en sí.

5.3.3.Generador de lenguaje

Dentro de este sistema se utilizará el generador de lenguaje para la generación del contenido del mensaje que nuestra cuenta va a publicar en Twitter. Para ello, nuestro sistema cuenta con varias funciones que permiten generar varios tipos de frases.

Como hemos dicho anteriormente, el generador de lenguaje será llamado cada vez que un usuario tenga que publicar contenido en Twitter. Para ello utilizaremos las librerías de simpleNLG para Java, que nos permiten generar contenido de una forma sencilla.

En primer lugar, antes de generar cualquier frase, simpleNLG necesita su propia configuración inicial. En primer lugar se lexicon, NLGFactory, and realiser que nos permitirá empezar a crear frases.

Una vez hemos creado estos tres objetos, SimpleNLG nos permite generar frases de dos formas distintas:

- La primera opción para generar una frase es la creación en un solo paso, es decir, se creará un NLGElement, que son las mínimas entidades para generar frases en la API, y se le pasará por parámetro la frase entera. Por tanto, El código para generar una frase de esta forma sería algo parecido a:

```
NLGElement Phrase = nlgFactory.createSentence("my dog is happy");
```

  De esta forma creamos una frase completa indicando el conjunto de toda ella. Ahora lo único que necesitamos es convertirla en un String, que nos permita insertarla como texto para el API de Twitter4j. Para ello realizar esto necesitamos un realiser, que comentamos previamente, que nos permite convertir un elemento del API en un String, de tal forma que el código quedaría de la siguiente forma:

```
String Tweet=realiser.realiseSentence(Phrase);
StatusUpdate status = new StatusUpdate(Tweet);
```

  Con estos comandos conseguimos crear una entidad de StatusUpdate que es la información que vamos a enviar mediante el API de Twitter4j. En el StatusUpdate se pueden añadir y cambiar la información adjunta al tweet. Algunos de los elementos de la información que podemos añadir es la geolocalización o los archivos adjuntos al tweet. Una vez ya hemos añadido o cambiado toda la información que deseamos del tweet lo

único que nos falta es enviar la información a Twitter para que se actualice nuestro estado actual. Para ello nos apoyaremos en el objeto twitter, que como hemos visto anteriormente, guarda la información de la cuenta. De tal forma que:

```
Status status2 = twitter.updateStatus(status);
```

Creamos un objeto Status que nos permitirá obtener la información sobre la ID de nuestro tweet, que se guardará en la base de datos (de la que hablaremos más adelante) que nos permitirá saber los tweets que se han enviado durante el experimento.

- La segunda opción es la generación de las frases mediante el uso de cláusulas. Estas cláusulas nos permiten modificar cada uno de los elementos de la frase por separado, es decir, podemos generar una frase mediante la introducción de sujeto, verbo y complementos por separado. Esto nos permite un mayor grado de adaptabilidad incluyendo cada uno de las partes de la frase por separado.

  El API de SimpleNLG consta de varios tipos de cláusulas dependiendo del tipo de frase que deseemos crear. En la siguiente tabla se muestran los tipos de cláusulas que se pueden crear y lo que incluyen.

Parte de la frase	Tipo de frase	Método
Subject	Noun Phrase	setSubject("")
Verb	Verb Phrase	setVerb("")
Object	Noun Phrase	setObject("")
IndirectObject	Noun Phrase	setIndirectObject("")
Complement	Preposition Phrase That-clause Adjective Phrase AdverbPhrase	addComplement("")
Modifier	That-clause Adjective Phrase Adverb Phrase	addModifier("")

Como vemos en la tabla existen diferentes tipos de frase. Esto se refiere al tipo de sintagma que es para que cuadre con la función que este debe realizar en la frase.

Una vez hemos generado cada uno de los sintagmas, estos se añadirán a un objeto SPhraseSpec que nos permite agregar todo, desde el sujeto a varios modificadores y complementos. Esto se realizará mediante las métodos que aparecen en la última columna de la tabla anterior.

Hecho esto, lo único que nos quedaría seria generar el String que nos permitiera añadir la información al tweet. Para ello se seguirán los mismos pasos que se han explicado previamente en el caso simple.

Viendo ambos tipos de generación de clases que nos permite *simpleNLG* se ha optado por el uso de la segundo debido a su modularidad y su facilidad de uso para nuestro caso especifico ya que nos permitirá generar un mayor número de frases posibles de una manera más sencilla.

Una vez hemos visto como se generan las cláusulas, vamos a ver los tipos de cláusulas que podemos generar mediante este procedimiento.

En primer lugar, viendo la gramática inglesa vemos que los tipos de oraciones simples que se pueden crear son los siguientes:

➢ TABLA 5:TIPOS DE FRASE

Identificador	Componentes
Frase 1	NP1 + V_be + ADV/TP
Frase 2	NP1 + V_be + ADJ
Frase 3	NP1 + V_be + NP1
Frase 4	NP1 + LV + ADJ
Frase 5	NP1 + LV + NP1
Frase 6	NP1 + V_int
Frase 7	NP1 + V_tr + NP2
Frase 8	NP1 + v_TR + NP2 + NP3
Frase 9	NP1 + V_tr + NP2 + ADJ
Frase 10	NP1 + V_tr + NP2 + NP2

En la tabla anterior se muestran todos los tipos sencillos de frase existentes en el inglés. En cada una de estas, cada una de las claves representa algo distinto.

- **NP:** Se trata de un sintagma nominal.
- **V_be:** Es el verbo to be.

- **ADV/TP:** Sintagma Adverbial de tiempo o lugar.
- **ADJ:** Un adjetivo calificativo.
- **LV:** Linking verb.
- **V_tr:** Verbo transitivo.
- **V_int:** Verbo intransitivo.

Una vez aclarados cada una de las claves, pese a que todos los tipos están implementados dentro del código del sistema únicamente se utilizaran las frases 1, 2, 4, 6 y 7. A continuación diremos las razones por las que hemos excluido los otros tipos de cláusulas.

En primer lugar, centrándonos en los tipos de frase 3 y 5, vemos como tiene dos sintagmas nominales con el mismo índice, el sintagma 1 en este caso. En este tipo de frase el segundo sintagma nominal hace referencia al primero, ya siendo nombrándolo otra vez o incluyendo algo más de información. Este segundo sintagma nominal realiza la función de complemento directo. Al haber una relación directa entre estas dos partes de la frase se ha decidido excluir este tipo para aumentar la probabilidad de que las frases generadas aleatoriamente tengan sentido.

Por otro lado, tenemos las frases 8, 9 y 10, las cuales tienen un mayor número de parámetros para formarlas. La razón de la exclusión es la misma que para los dos tipos anteriores, aumentar la probabilidad de generar una frase con sentido. Al tener un mayor número de componentes y al elegir los componentes de una forma aleatoria la probabilidad de que estos tengan sentido entre sí es muy baja.

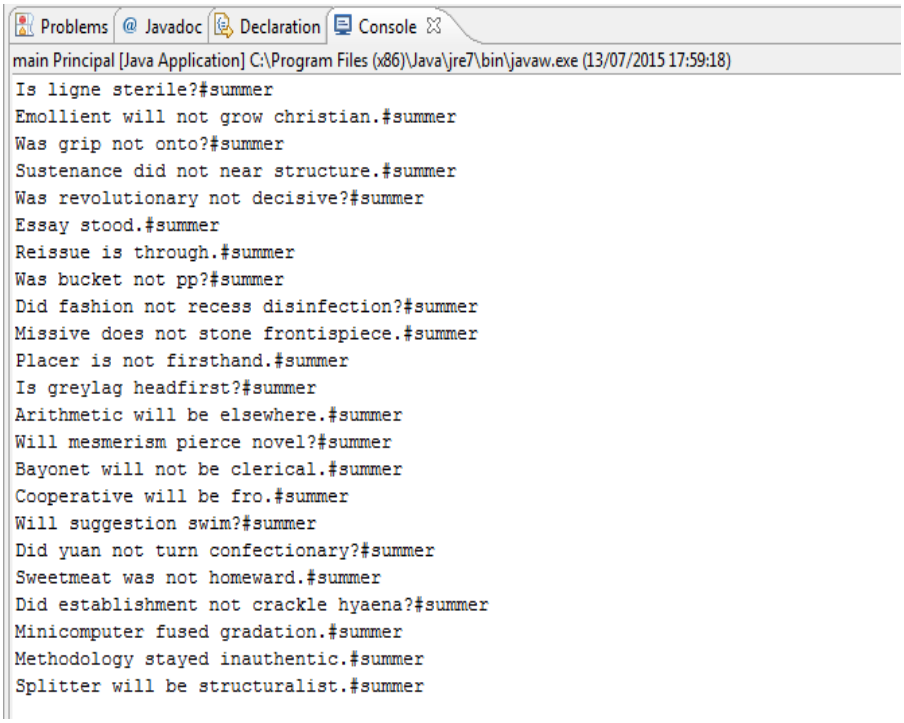
Además de los tipos de cláusulas que podemos generar, tenemos más grados de libertad para generar distintos tipos. Entre ellos tenemos:

- **Tiempo verbal:** La primera modificación que podemos realizar en la frase es la modificación del tiempo verbal. Entre las posibles opciones que podemos seleccionar son: pasado, presente y futuro.
- **Frases interrogativas:** la segunda modificación que podemos realizar es la generación de frases interrogativas. El API de simpleNLG nos permite la generación de frases interrogativas, modificando el orden de los elementos de las cláusulas para que este se corresponda con el orden correcto de la frase.

- **Frases negativas:** Por último, simpleNLG nos permite la modificación entre cláusulas afirmativas y negativas incluyendo el verbo auxiliar si fuera necesario para su correcta gramática.

Estas características comentadas anteriormente se pueden realizar de manera concurrente, es decir, podríamos generar una frase futura, interrogativa y negativa. Esto nos permite una gran capacidad de modificación de las frases y aumenta el número posible de frases generables.

Algunos ejemplos de las frases generadas se pueden ver en la siguiente imagen, donde los mensajes enviados por cada uno de los usuarios se recogen en la consola.



➢ ILUSTRACIÓN 21:EJEMPLOS DE FRASES

Como vemos, hay combinaciones de todos los tiempos verbales con frases tanto afirmativas como negativas y algunas preguntas. Esto, como hemos dicho anteriormente, hace que nuestro sistema tenga una gran variedad de posibilidades para generar frases.

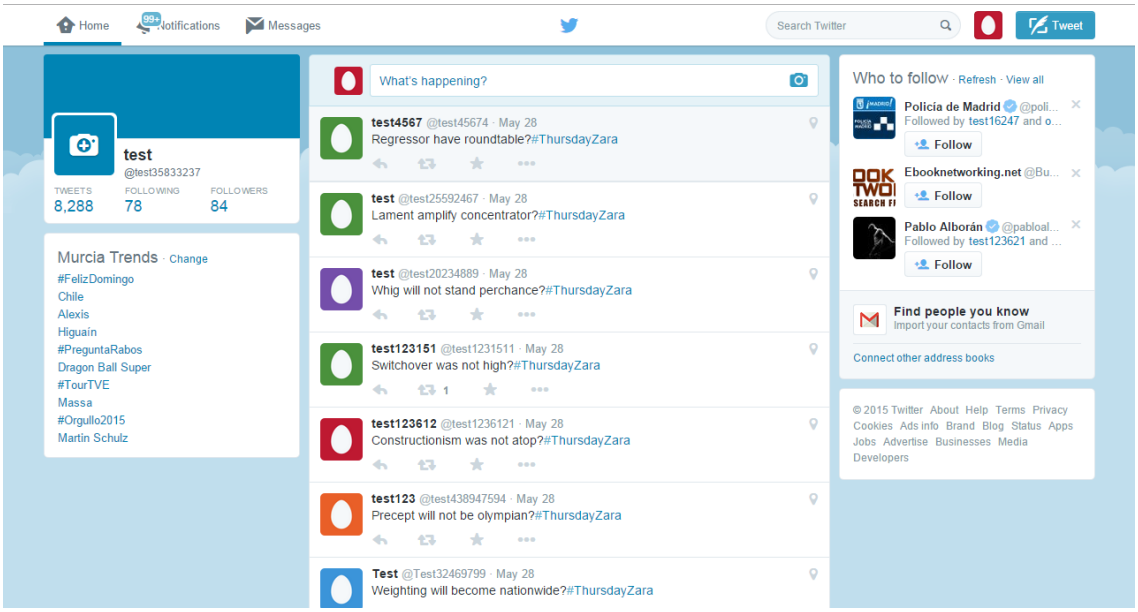
Ejemplo de ejecución del sistema de usuarios

Una vez hemos visto las labores que desempeña este sistema dentro de la herramienta vamos a ver cada uno de los TimeLine de algunos de los usuarios

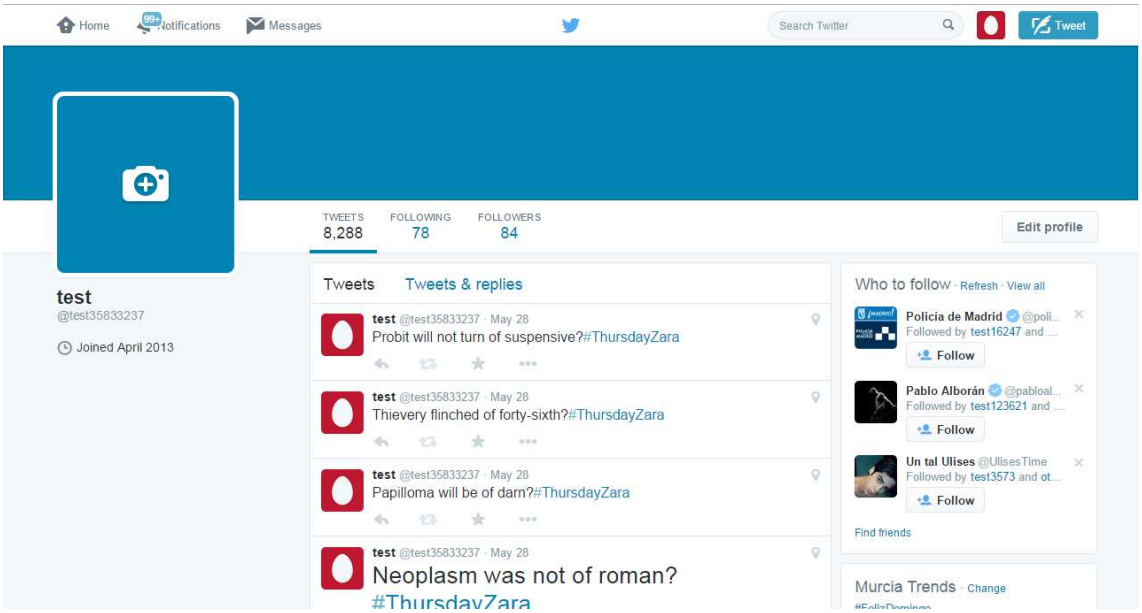
activos durante el experimento para verificar que estos funcionan correctamente.

En primer lugar mostraremos la figura del primer usuario, que es el que realizará la labor de emitir un mensaje en Twitter. Posteriormente veremos cómo este mensaje llega a todos sus seguidores.

Para mostrar que la herramienta funciona correctamente vamos a realizar la comparación entre el antes y el después del Tweet de ambas cuentas.

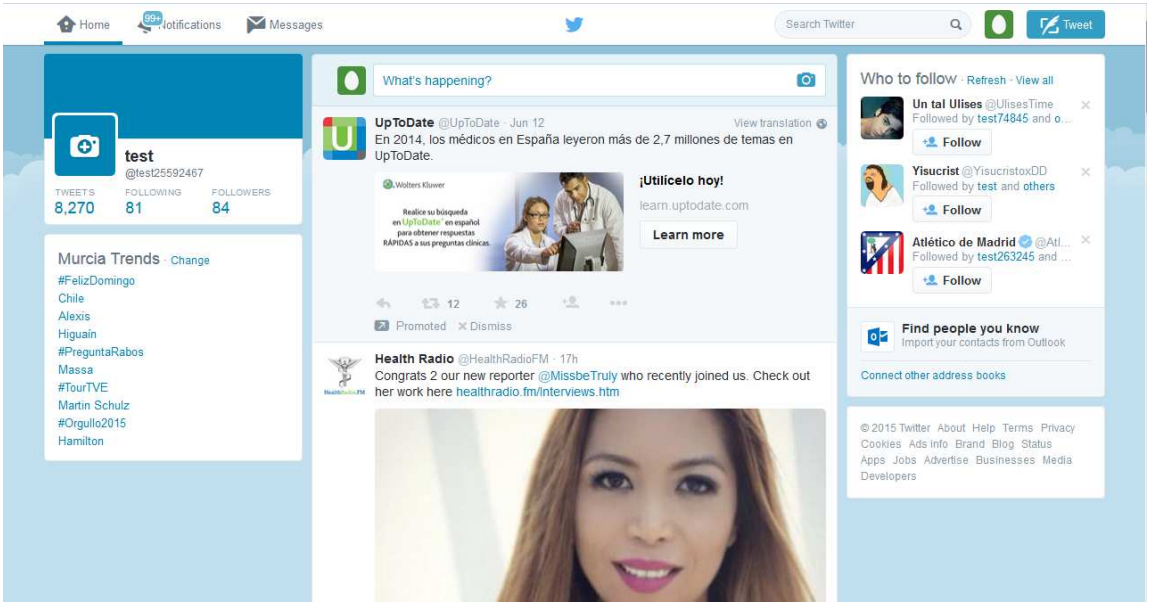


➢ ILUSTRACIÓN 22:TIMELINE USUARIO 1



➢ ILUSTRACIÓN 23:TWEETS ENVIADOS ANTES POR EL USUARIO 1

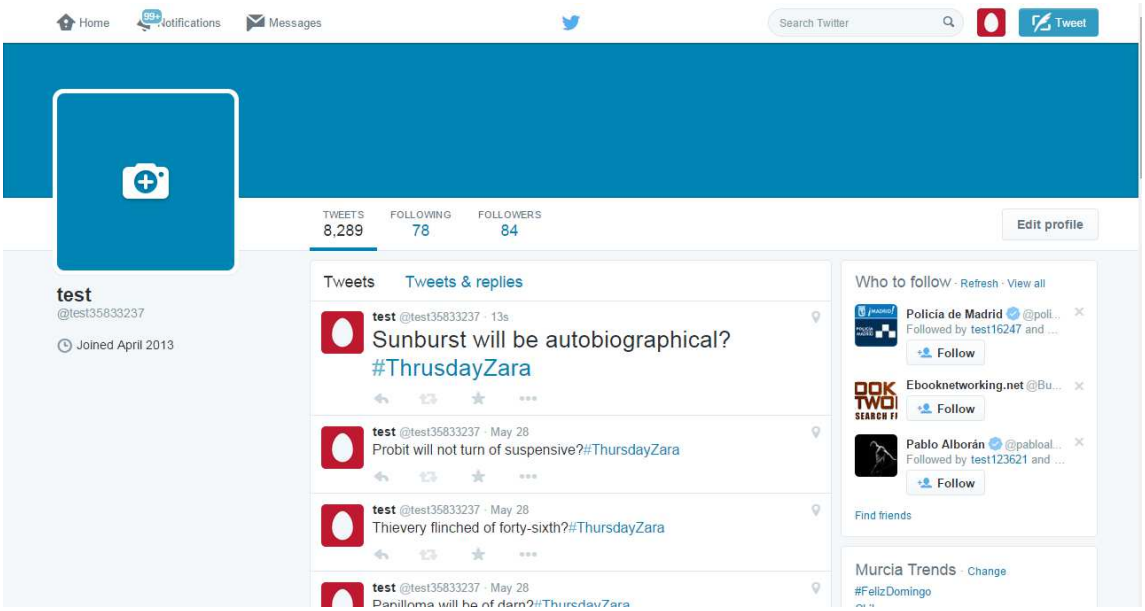
En estas dos primeras imágenes se muestran el TimeLine y la lista de Tweets que han sido enviados con esta cuenta. Como vemos en ambos casos, el último mensaje realizado data del 28 de Mayo.



➢ ILUSTRACIÓN 24:TIMELINE ANTES DEL USUARIO 2

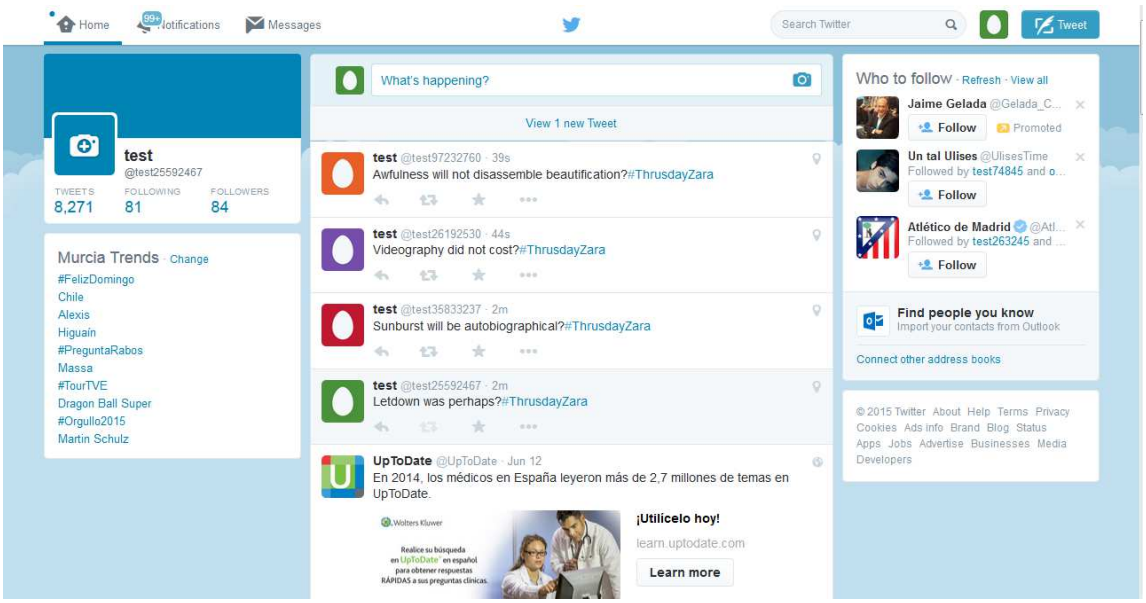
En esta tercera imagen se muestran los Tweets recibidos por un segundo usuario que está dentro del sistema. Como vemos el último mensaje recibido data de hace 17 horas.

Una vez hemos visto la situación de reposo de ambas cuentas vamos a ver lo que sucede una vez ejecutamos el código.



➢ ILUSTRACIÓN 25:TWEETS ENVIADOS DESPUÉS POR EL USUARIO 1

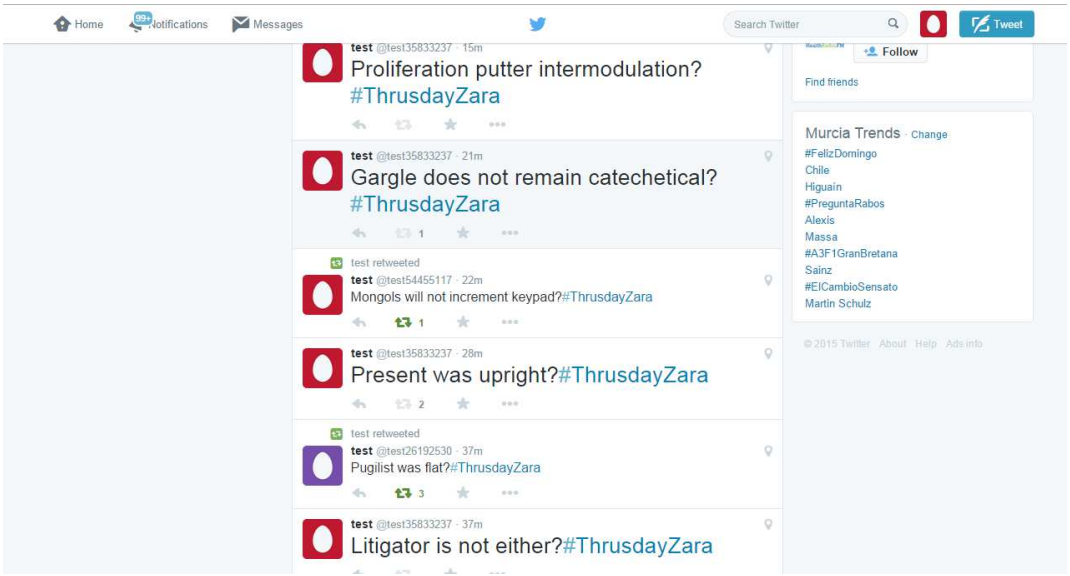
En esta imagen se muestra la pagina donde se muestran los Tweets enviados por los usuarios. Como se puede observar en la parte superior de la ilustración, el último mensaje enviado data de 13 segundos antes de que se tomará la imagen.



➢ ILUSTRACIÓN 26:TIMELINE DESPUÉS USUARIO 2

En la ilustración anterior se muestra como el mensaje enviado por la otra cuenta, junto a las de otras muchas, se muestran en el TimeLine de otros usuarios virtuales que participan dentro del experimento.

Por último, se va a mostrar una imagen posterior del TimeLine de este segundo usuario donde se ve y aprecia como algunos de los Tweets que envía son retweets. Esta imagen se ha tomado algo de tiempo después que las previamente analizadas ya que los usuarios necesitan un tiempo hasta que estos realizan retweets debido a su probabilidad que fue comentada anteriormente.



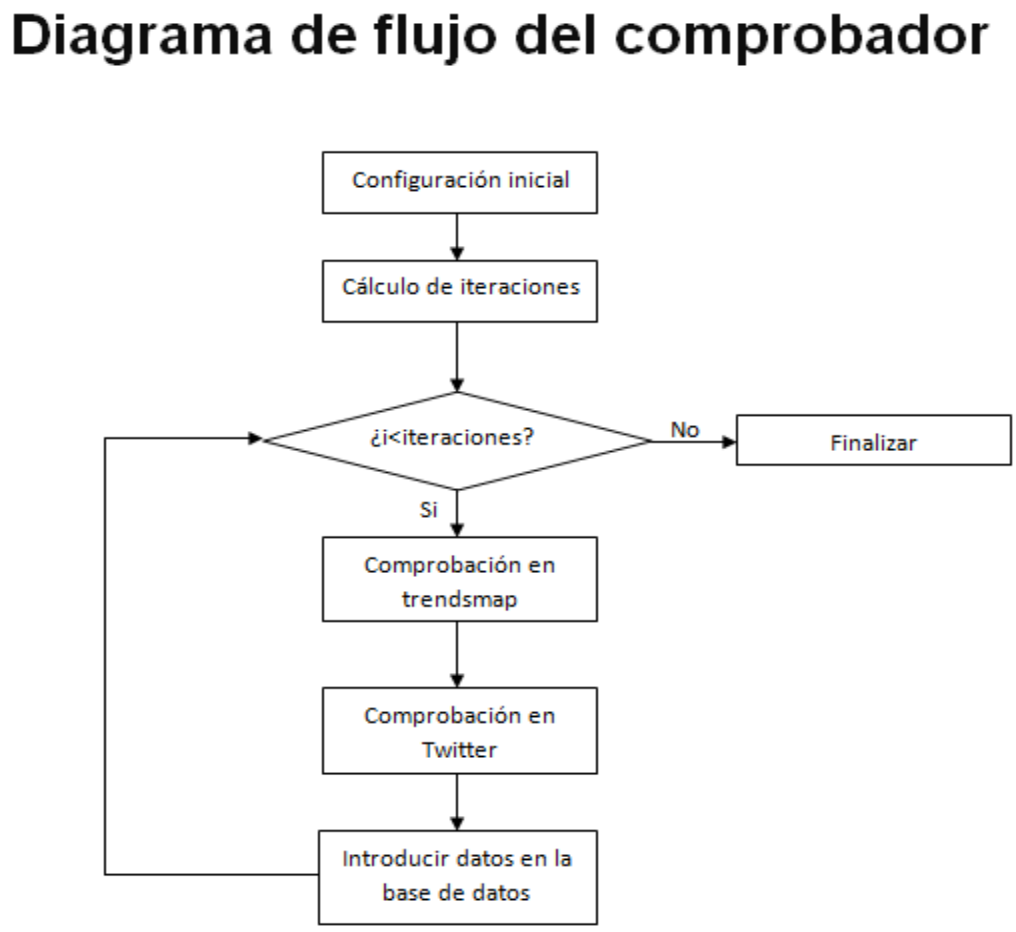
➢ ILUSTRACIÓN 27:RETWEETS EN LAS CUENTAS

5.4. Comprobador

Por último, tenemos el sistema que comprueba si nuestra ejecución ha tenido éxito o , sin embargo, no ha logrado su objetivo. Antes de pasar a ver el diagrama de flujo de este sistema, vamos a comentar la página de trendsmap.com. Esta página es una página externa que nos ofrece información de las tendencias más habladas del momento. Esta nos ofrece información de si nuestro hashtag está entre las tendencias más habladas del momento en la geolocalización elegida. Posteriormente se explicara cómo se realiza la comprobación tanto en esta página como en Twitter.

Para ver el conjunto de tareas que se realizarán en este sistema vamos a analizarlo como lo hemos hecho previamente. En primer lugar mostraremos el diagrama de flujo del sistema para después explicar el funcionamiento paso a paso.

Por tanto, aquí tenemos el diagrama de flujo del comprobador.



➢ ILUSTRACIÓN 28:DIAGRAMA DE FLUJO DEL COMPROBADOR

A continuación explicaremos en detalle cada uno de los pasos que seguirá nuestro sistema de comprobación.

- En primer lugar se realiza la configuración inicial del sistema. Esto consiste en:
    - Obtención del tiempo total de ejecución del sistema completo.
    - Obtención del hashtag que se está utilizando en el experimento.
    - Selección del woeid según la geolocalización en la que se está realizando el experimento.
    - Selección de la página de trendsmap a la que se deberá acceder. Según la geolocalización esta página será una u otra [26].
    - Obtención de la geolocalización que se está utilizando en el experimento.
    - Obtención del número de usuarios que se está utilizando en el experimento.

- Toda esta información será necesaria para la correcta documentación de los experimentos realizados.

- Tras realizar la configuración inicial del sistema se realizará el cálculo del número de iteraciones que serán necesarias para que este sistema vaya comprobando a lo largo de la vida del experimento. Como deseamos que se compruebe cada 30 minutos desde su inicio el número de iteraciones será:

$$iteraciones = \frac{TiempoTotal}{30}$$

  Esta variable será la que nos permitirá saber cuando este sistema debe acabar su ejecución.

- El siguiente paso que deberemos cumplir es la comprobación en la página correspondiente de trendsmap. Como hemos dicho previamente en la configuración inicial del sistema, según la geolocalización en la que se esté realizando el experimento la página variará.

- La comprobación de esta página consiste en la adquisición mediante el protocolo HTTP de la página de trendsmap. Una vez la hemos obtenido buscaremos dentro de su contenido los temas más hablados del momento y, en función de si el hashtag que estamos utilizando se encuentra dentro de esta lista, sabremos si es uno de los más utilizados en el momento.

- La otra comprobación que debemos realizar es la comprobación dentro de Twitter. Para ello utilizaremos una de las cuentas utilizadas dentro del experimento que, mediante el API de Twitter4j, podremos obtener los temas más hablados en la zona que estamos publicando. Para ello, lo único que necesitaremos será el woeid de la zona en la que estamos publicando.

- Por último, una vez realizadas las comprobaciones en ambos lugares, se procederá al almacenamiento de la información dentro de la base de datos para su posterior análisis.

Este sistema es el único que utiliza la base de datos para guardar información dentro de ella, ya que es el único que genera información en el experimento. Tanto el planificador como el sistema de usuarios utilizan la base de datos para obtener información, ya sea de cuentas de Twitter o la utilizan para obtener palabras que se utilizarán en el generador de lenguaje.

Como se puede ver, este sistema se encarga de comprobar el efecto de cada uno de los experimento en dos lugares distintos, dentro de la propia red social de Twitter y en la página de trendsmap, que nos permite ver en directo cuales son los temas sobre los que más se esta tweeteando en un determinado momento en una determinada geolocalización. Con este sistema sabremos cuando el número de usuarios utilizado para un determinado momento son suficientes para lograr que estos causen repercusión en la red.

La comprobación en cada una de estos dos lugares en bastante sencilla.

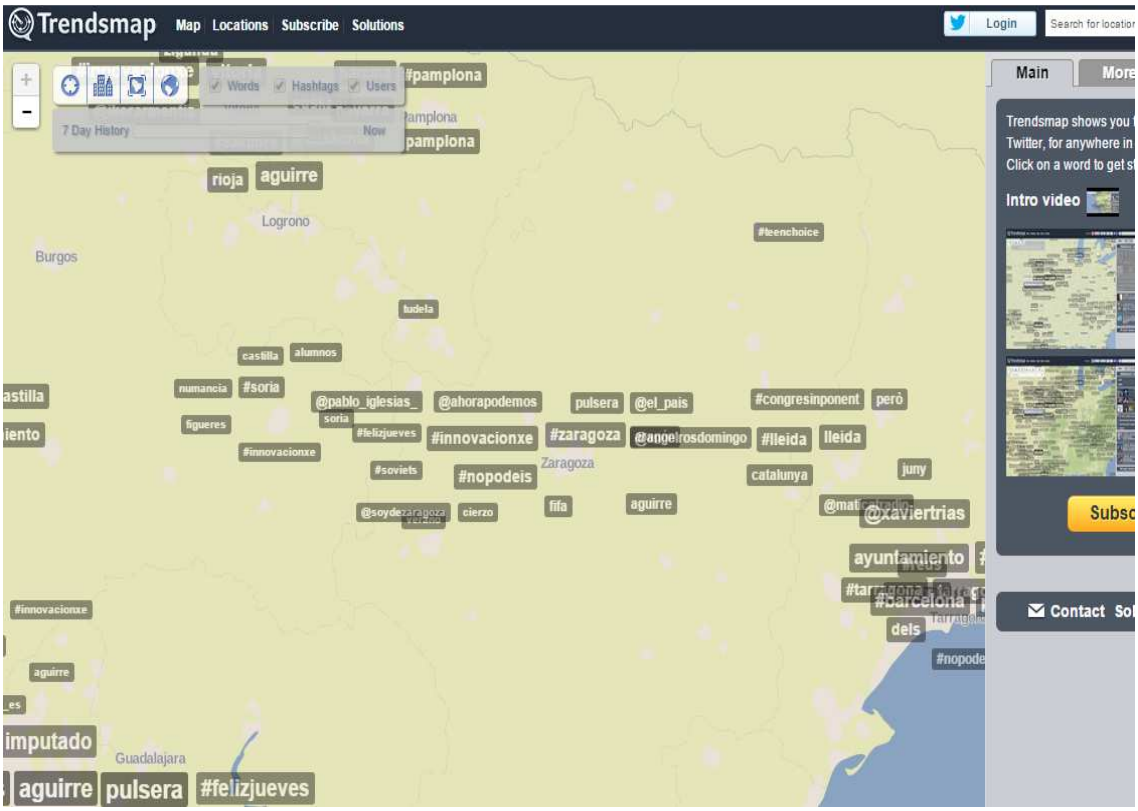
Twitter

Para realizar la comprobación en Twitter se realizará mediante las librerías que nos proporciona Twitter4j. Para ello, utilizaremos una cuenta de usuario de nuestra base de datos para crear una entidad *twitter.* Una vez hemos creado esta entidad, comprobando la localización en la que se está publicando, cogeremos el woeid correspondiente a la geolocalización utilizada para coger los Trending Topic del momento. Una vez obtenidos, se procederá a la comparación de cada uno de los elementos de esta lista con el hashtag utilizado. En cualquiera de los casos, el resultado obtenido de la comprobación con los Trending Topic del momento se guardará en la base de datos para futuros análisis de los resultados.

Trendsmap

En cuanto a la comprobación en la página trendsmap, como se ha comentado previamente se realizará mediante el análisis de el contenido HTML de la página. Para ver el funcionamiento de esta página vamos a ver la información que esta nos proporciona lo largo de un experimento.

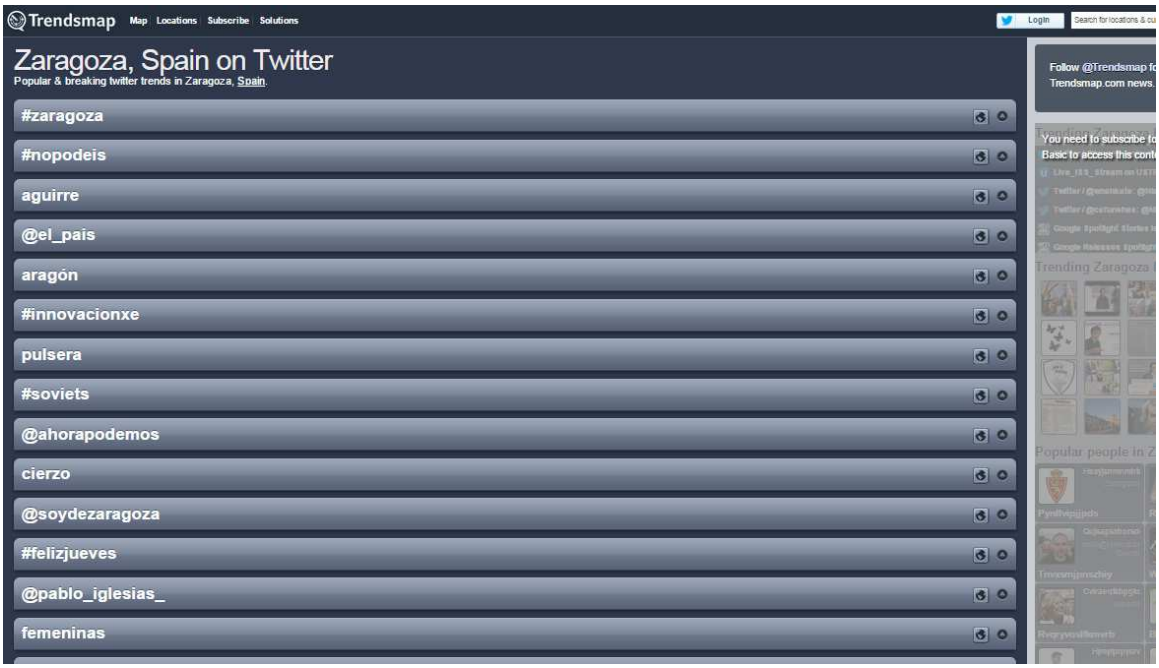
Se realizará la prueba con un experimento para ver como varía la información de la página trendsmap. En primer lugar vamos a ver la página antes de ejecutar nuestro programa. La geolocalización elegida para el experimento ha sido Zaragoza y el número de usuarios que intervinieron en este fue de 80.La duración del experimento es de una hora, que es el tiempo que se utiliza generalmente para todos los experimentos. El hashtag que se ha utilizado es #thursdayZara.



➢ ILUSTRACIÓN 29:TRENDSMAP ANTES DEL EXPERIMENTO 1

En esta primera imagen vemos la página inicial de trendsmap, donde aparecen todos los temas más hablados del momento en el mundo. En particular estamos viendo la zona de Zaragoza y los temas más hablados del momento en esta zona. Como vemos, estos temas tienen una posición determinada por el lugar de donde provienen los tweets que lo utilizan. Es por tanto fácil distinguir los

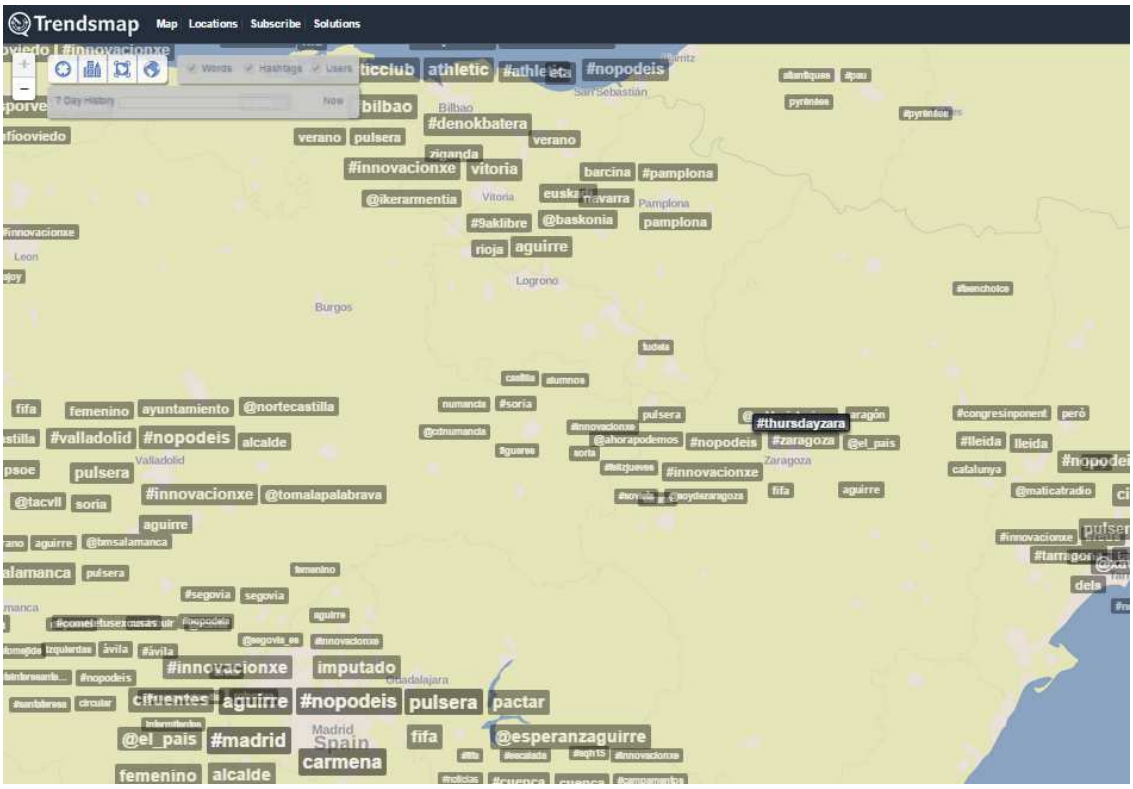
temas más hablados en cada una de las regiones que hemos seleccionado para realizar los experimentos. Hay que tener en cuenta que el tamaño de cada una de las etiquetas en el mapa indica la importancia que ese tema tiene, es decir, cuanto más grande sea la etiqueta más se está hablando de eso.



➢ ILUSTRACIÓN 30:TRENDSMAP ANTES DEL EXPERIMENTO 2

En esta segunda imagen vemos una lista de los temas más hablados en cada una de las geolocalizaciones. Cada una de las geolocalizaciones tiene una página correspondiente donde se puede ver esta lista. Además, esta web nos ofrece una lista por orden de importancia en el momento en que la consultamos, es decir, los temas más hablados están colocados al inicio de esta y estos se van colocando en orden decreciente. Debido a que la información en esta segunda web esta mejor ordenada y simplificada, ya que solo tenemos los temas más hablados de esta región, a partir de esta segunda página obtendremos toda la información necesaria para nuestro experimento. Al igual que hicimos con el caso de la comprobación en twitter, aquí utilizaremos la información del fichero html para obtener una lista con los temas más hablados. Una vez hemos adquirido esta lista, lo único que tenemos que hacer es comparar con el hashtag que nosotros hemos utilizado a lo largo del experimento.

Una vez hemos visto la página en estado de reposo y hemos explicado brevemente como se realizará la comprobación de si nuestro hashtag aparece en esta web, vamos a ver las mismas imágenes una vez se ha ejecutado el experimento.



➢ ILUSTRACIÓN 31:TRENDSMAP DESPUÉS DEL EXPERIMENTO 1

En primer lugar vemos la imagen de la página inicial de trendsmap, al igual que hicimos anteriormente. Como vemos, el hashtag que hemos utilizado durante el experimento, *#thursdayZara* se ha colocado entre los temas más hablados de la zona.

➢ ILUSTRACIÓN 32:TRENDSMAP DESPUÉS DEL EXPERIMENTO 2

Como vemos en esta segunda imagen, nuestro hashtag se ha colocado en el tercer lugar entre los temas más hablados del momento.

Al ver ambas imágenes, podemos decir que nuestro experimento ha tenido éxito, ya que en ambas aparece nuestro mensaje como uno de los más populares. Cabe destacar que, en este caso hemos utilizado un gran número de usuarios para esta localización, como veremos más adelante con los datos que hemos obtenido al ejecutar nuestro proyecto. Aun así, vemos que aparecer entre los temas más hablados no es una tarea imposible en esta web especializada.

Por último, destacar al estar comprobando en la web de cada una de las regiones especializadas donde aparece la lista de temas más populares, en algunos de los casos el hashtag utilizado en los experimentos no aparecerá en la página principal debido a que en esta aparece un menor número de entradas, ya que en los experimento lo que se intenta es saber el número mínimo de usuarios que debemos involucrar para conseguir aparecer en la web de cada una de las regiones, y no importa en el puesto en el que aparezca.

5.5. Base de datos

Como hemos dicho previamente, la base de datos es una parte fundamental de nuestro sistema completo, ya que cada uno de los pequeños sistemas necesitan de su uso.

La base de datos está diseñada en mySQL y consta de cuatro tablas distintas, que permiten el funcionamiento correcto del sistema completo. A continuación mostraremos cada una de las tablas de las que consta la base de datos con cada uno de sus atributos y las funciones que logran solventar.

Account

La primera tabla que estudiaremos en profundidad será la de *account*. Está primera tabla nos permite el almacenamiento de la información de todas las cuentas de Twitter que podremos utilizar. En la siguiente tabla mostramos todos los atributos de esta.

➢ TABLA 6: TABLA ACCOUNT

Nombre	Tipo
Name	Varchar 20
Password	Varchar 20
E-mail	Varchar 30
API-key	Varchar 40
API-secret	Varchar 50
AccessToken	Varchar 60
AccessTokenSecret	Varchar 50
used	int

Como vemos en la tabla anterior, *account* cuenta con 8 atributos distintos que sirven para:

- **Name:** es el nombre de referencia que la cuenta utiliza en la red social de Twitter.
- **Password:** Se guarda la información de la contraseña de usuario. El password unido con el e-mail o el nombre de la cuenta nos permitiría iniciar sesión en Twitter.

- **E-mail:** Como bien indica el nombre es la cuenta de correo asociada a la cuenta de Twitter.

- **API-key:** Esta es la clave de acceso pública de nuestra aplicación. Esta información junto al API-secret indica la información de la aplicación que se utiliza para comunicarse con Twitter.

- **API-secret:** Esta es la clave de acceso privada de nuestra aplicación. Como hemos comentado previamente, este atributo permite a la aplicación comunicarse con el API de Twitter.

- **AccessToken:** Es la clave de acceso pública de la cuenta. Esta nos permite, junto al AccessTokenSecret publicar u obtener información con nuestra cuenta de usuario.

- **AccessTokenSecret:** Es la clave de acceso privada de la cuenta. Esta nos permite, junto al AccessToken publicar u obtener información con nuestra cuenta de usuario.

- **Used:** Este atributo guarda la información de si la cuenta esta actualmente en uso. Este atributo es el que se comprobará cada vez que se intente añadir una nueva cuenta dentro del experimento.

Esta tabla nos permite tener la información de todas las cuentas ordenada, permitiendo, con el uso del atributo used, que no se utilicen varias veces algunas cuentas, lo cual pondría en riesgo las políticas de seguridad de Twitter sobre spam, ya que al utilizar varias veces la misma cuenta se multiplicaría el número de tweets que esta cuenta pública.

Word

La siguiente tabla a estudiar es la tabla *word.* Esta tabla es la tabla utilizada por el sistema generador de lenguaje ya que en esta tabla se encuentran todas la palabras que se van a utilizar en las posibles frases que se generen. Mirando más detenidamente esta tabla encontramos que los atributos son:

➢ TABLA 7: TABLA WORD

Nombre	Tipo
Word	Varchar 30
Identificador	Varchar 4
ID	Int

Como vemos en la tabla anterior, *word* cuenta con solo tres atributos:

- **Word:** En este atributo se guarda la palabra en sí. Es el lugar del que el sistema generador de lenguaje obtendrá cada una de las palabras.

- **Identificador:** El identificador nos permite diferenciar el tipo de palabra que es. Es decir, si la palabra se trata de un sustantivo, un verbo, un adjetivo... Este atributo nos permite buscar las palabras de una forma más eficiente.

- **ID:** Por último, tenemos la ID, que se trata de un número que identifica a cada una de las palabras de cada uno de los identificadores. Esta ID nos permite obtener cada una de las palabras de forma aleatoria.

Como hemos dicho previamente, esta tabla es la principal en el generador de lenguaje ya que proporciona todas las palabras necesarias para componer cada una de las frases. Para ello, contamos con hasta 20 identificadores distintos, que otorgan mayor variedad a la posibilidad de generación de nuevas frases. De los 20 tipos de palabras dentro de la base de datos, las estadísticas de cada uno de los tipos más utilizados son los siguientes:

➢ TABLA 8: ESTADÍSTICAS DE WORD

Identificador	Número de entradas
NN (Sustantivos)	21992
RB (Adverbio)	378
VB (Verbos transitivos)	6272
JJ (Adjetivos)	8361
IN (Preposiciones)	101
DT (Pronombre)	121
MD (Verbos modales)	19
VBL (Verbos copulativos)	14
VBI (Verbos intransitivos)	36

Este gran número de entradas en cada una de los distintos identificadores nos permite poder generar una gran variedad de posibles frases, que uniéndolo a la capacidad de personalización de las mismas que nos proporciona *simpleNLG* las posibles combinaciones son casi imposibles que se puedan repetir, incluso en diferentes experimentos.

TestInfo

La tabla *testinfo* nos proporciona la información general de cada uno de los experimentos realizados. Los atributos de esta tabla son:

➢ TABLA 9:TABLA TESTINFO

Nombre	Tipo
ID	Varchar 20
Topic	Varchar 20
Geolocalizacion	Varchar 20
Users	Int
Distribution	Varchar 15
Time	Int

Vamos a describir más detenidamente la función de cada uno de ellos:

- **ID:** Es el identificador del experimento. Este atributo nos permite la identificación univoca de los experimentos.
- **Topic:** El topic guarda la palabra que se utilizará como hashtag en el experimento.
- **Geolocalización:** Lugar donde se realiza el experimento.
- **Users:** Guarda el número de usuarios totales que participan dentro del experimento.
- **Distribution:** Es el tipo de distribución que se utilizará en el experimento. En nuestro caso solo tenemos tres: lineal, exponencial e intensiva.
- **Time:** Por último tenemos el tiempo total de ejecución. Este tiempo determina además del tiempo total de ejecución el número de entradas que habrá en la tabla check.

Esta tabla guardará la información básica del experimento en sus atributos. La entrada de esta tabla se genera al inicio de cada uno de los experimentos.

Checks

Por último, la tabla *checks* guarda la información generada por el comprobador en cada una de sus iteraciones. En este se guarda la información generada por cada uno de los experimentos realizados. Los atributos de esta tabla son:

➢ TABLA 10:TABLA TEST

Nombre	Tipo
ID	Varchar 20
CheckTime	Int
Twitter	Int
TrendsMap	Int

Como vemos, consta de cuatro campos que se utilizan para:

- **ID:** Es el campo de la ID del experimento. Habrá múltiples referencias de varias tuplas con la misma ID, esto es debido a que se guardan varias instancias de checks por cada ejecución en función de la duración del mismo.
- **CheckTime:** Es el tiempo relativo al inicio del experimento en el que se realizo la comprobación.
- **Twitter:** Es la comprobación en el propio Twitter.
- **Trendsmap:** Es la comprobación en la página de trendsmap.

Esta tabla nos servirá para guardar la información de cada uno de los experimentos realizados, guardando si el experimento ha conseguido aparecer entre los temas más populares de la página trendsmap y si el hashtag ha aparecido entre los TTs de Twitter en la geolocalización correspondiente. El proceso de comprobación se realiza cada 30 minutos, comprobando en ambos lugares si se ha conseguido aparecer como uno de los temas más populares en el momento. Es por ello que, dependiendo de la duración del experimento habrá un número mayor o menor de entradas en la tabla *checks*.

6.Datos obtenidos

Una vez se ha explicado el funcionamiento como la implementación del sistema que va a realizar los experimentos vamos a explicar en qué van a consistir estos experimentos.

Los tipos de experimentos que se van a realizar van a ser dos, que consisten en los siguiente:

- El primer experimento que se realizará será el análisis del número de cuentas que son necesarias para aparecer en la página trendsmap. Esta página, como se ha comentado previamente, nos ofrece información sobre los temas más hablados en el momento en una determinada región geográfica. Con esto se pretende averiguar cuál es el número mínimo de usuarios necesarios para alcanzar ser uno de los temas más hablados encada una de las zonas preestablecidas. Esta información será necesaria para realizar el segundo tipo de experimentos además de darnos información sobre el esfuerzo necesario para aparecer como tendencia de moda.

- El segundo experimento que se realizará será el test intensivo de nuestro programa en la zona que necesite menor número de usuarios para alcanzar ser uno de los temas más hablados. Al utilizar la zona con menor número de usuarios necesarios estamos maximizando las posibilidades de que nuestro hashtag se convierta en Trending Topic en Twitter.

Una vez se han diferenciado los dos tipos de experimentos que se realizarán vamos a ver como se realizaran cada uno de ellos más detenidamente.

6.1. Usuarios en cada geolocalización

En el primer tipo de experimentos que hemos comentado, se basa en la ejecución repetida de nuestro sistema con el fin de encontrar el número mínimo de usuarios necesarios para que el tema que se ha escogido como hashtag logré aparecer entre los temas más hablados del momento. Es por ello que, este tipo de experimentos tienen una duración corta, de una hora exactamente, en la cual las distintas cuentas publicarán cumpliendo una distribución exponencial. En caso de que el experimento en cuestión de resultados, se pasará a añadir el número de usuarios necesarios en este a la base de datos. En caso de que el experimento no logré su fin, se realizará una pausa, generalmente 1 o 2 horas, en las cuales no se publicará, con el fin de que los distintos experimentos sean independientes entre sí, y los datos obtenidos no se vean alterados por ejecuciones anteriores.

Otra de las características de este tipo de experimento es que el número de usuarios crecerá en caso de que el experimento anterior no tuviera efecto. De esta forma lograremos , aproximadamente, saber cuál es el número de usuarios necesarios. El crecimiento del número de usuarios será en múltiplos de 10, que nos permitirá de una forma rápida obtener el número de usuarios necesarios en cada una de las zonas.

Por último, la distribución elegida de todas las implementadas en el sistema es la distribución exponencial. Utilizaremos esta distribución ya que es la distribución con la que normalmente se generan los Trending Topic en Twitter, y queremos que el experimento se parezca lo máximo posible a una situación real.

Una vez hemos comentado el procedimiento mediante el cual vamos a realizar los experimentos, vamos a ver los resultados obtenidos en cada una de las geolocalizaciones.

➢ TABLA 11:DATOS DE LOS EXPERIMENTOS REALIZADOS

Experimento / Localización	#1	#2	#3	#4	#5	Media	Desviación típica
Murcia	20	30	30	20	20	24	5.48
Málaga	30	30	20	50	40	34	11.40
Sevilla	40	40	40	30	40	38	4.47
Valencia	50	40	40	30	40	40	7.07
Zaragoza	20	20	10	10	10	14	5.48
Bilbao	30	20	20	20	20	22	4.47
Barcelona	70	60	80	70	70	70	7.07
Madrid	-	-	-	-	-	-	-
Alicante	40	30	20	20	40	30	10
Córdoba	20	10	10	10	10	12	4.47
Valladolid	30	20	20	30	40	28	8.37
Vigo	20	10	10	20	20	16	5.48
Gijón	10	20	20	20	20	18	4.47
La Coruña	20	30	30	20	30	26	5.48
Vitoria	10	20	10	20	10	14	5.48
Granada	40	50	30	40	30	38	8.37
Oviedo	20	20	30	30	30	26	4.47

En la tabla anterior se muestran todos los experimentos realizados en todas las geolocalizaciones que el programa tiene predefinidos. Como vemos en los datos, el número de cuentas necesarias para que el experimento logrará aparecer en la página de trendsmap varía en función de de cada uno de ellos. Esto es debido a varios factores:

- El primer factor a tener en cuenta es la hora del experimento. En función de la hora que se esté ejecutando el experimento el número de usuarios activos en Twitter variará, siendo menor el número de usuarios en horario nocturno.

- El segundo factor es el día en el que se está publicando. Dependiendo del día que se esté realizando el experimento el número de usuarios activos varía siendo, generalmente, algo más elevado durante los fines de semana. Vinculado con el día de publicación también hay que tener en cuenta si se trata de una época donde haya una festividad importante (como por ejemplo semana santa o Navidad) que hace que el número de usuarios fluctúe entre zonas de acción. Este flujo puede ser tanto positivo, como puede ocurrir en las zonas de Andalucía en Semana Santa, o negativo, como lo es en Madrid en la misma época del año.

- El último factor a tener en cuenta son las noticias de última hora. Es decir, cualquier noticia de última hora puede afectar negativamente al experimento ya que este evento causará que las personas se vuelvan más activas en Twitter.

Teniendo en cuenta todos estos factores vemos, que a pesar de todo, el valor de fluctuación entre las distintas ejecuciones no es muy grande, ya que la máxima variación es de 20 usuarios.

Otro peculiaridad que podemos observar en la tabla es que en Madrid solo se ha conseguido que uno de los experimentos realizados logrará su objetivo. Esto es debido a que el número máximo de usuarios que se disponen es de 80, por lo que no se ha podido continuar la ejecución de experimento con mayor número de usuarios. Como se ha dicho previamente, el número de usuarios activos en la red durante el periodo nocturno es menor que durante el diurno, y es cuando uno de los experimento logró conseguir aparecer en la página.

Una vez hemos analizado los datos de los experimento realizados vamos a hacer una comparación con la población de cada una de las geolocalizaciones utilizadas. En este primera comparación, utilizaremos la población total de la ciudad que se está utilizando como experimento.

Para la obtención de los datos de población utilizaremos la página web del instituto nacional de estadísticas, que nos ofrece una información detallada de los datos de población tanto de los municipios como de las comunidades autónomas y provincias[27].

Los resultados obtenidos al comparar los datos obtenidos en el experimento con la población del municipio obtenemos lo siguiente:

➢ TABLA 12:POBLACION CIUDAD/USUARIOS

Localización	Población	Usuarios	Población/usuarios
Madrid	3.165.235	-	-
Barcelona	1.602.386	70	22891.23
Valencia	786.424	40	19660.6
Sevilla	696.676	38	18333.58
Zaragoza	665.058	14	47504.14
Málaga	566.913	34	16673.91
Murcia	439.712	24	18321.33
Bilbao	346.574	22	15753.36
Alicante	332.067	30	11068.9
Córdoba	328.041	12	27.336.75
Valladolid	306.830	28	10958.21
Vigo	294.997	16	18437.31
Gijón	275.735	18	15318.61
La Coruña	244.810	26	9415.77
Vitoria	242.082	14	17291.57
Granada	237.540	38	6251.05
Oviedo	223.765	26	8606.34

Como se puede ver en la tabla anterior, en la última columna donde se hace una comparación directa entre la población y el número de usuarios necesarios, se puede apreciar como este cálculo fluctúa bastante. En algunos casos, como el de Zaragoza, la relación es más del doble que el de todas las demás regiones.

Esto nos hace pensar que debe haber algo más que afecte a la hora de que la gente sea activa en la red social de Twitter.

Es por ello, que vamos a comparar los datos previamente generados con la población de cada una de las localidades elegidas comprendida entre los 14 y los 54 años que es la gente que es más activa dentro de las redes sociales.

➢ TABLA 13:POBLACIÓN CIUDAD ACTIVA/USUARIOS

Localización	Población	Usuarios	Población/usuarios
Madrid	2175995	-	-
Barcelona	1071112	70	15301.6
Valencia	537585	40	13439.62
Sevilla	490241	38	12907.08
Zaragoza	451466	14	32247.57
Málaga	409978	34	12058.18
Murcia	331020	24	13792.5
Bilbao	220993	22	10045.14
Alicante	235447	30	7848.23
Córdoba	233045	12	19420.42
Valladolid	193249	28	6701.75
Vigo	198024	16	12376.5
Gijón	170631	18	9479.5
La Coruña	156947	26	6036.42
Vitoria	165215	14	11801.07
Granada	160492	38	4223.47
Oviedo	146587	26	5637.96

Al comparar la población de entre 14 y 54 años lo que intentamos es comparar la población activa en las redes sociales con el número de usuarios que fueron necesarios en cada uno de los experimentos. De esta forma, como se puede apreciar, los datos obtenidos se vuelven algo más compactos, aunque sigue habiendo singularidades, como es el caso de Zaragoza, donde la relación entre la población y el número de usuarios sigue siendo más alta que los de las demás ciudades.

Como hasta el momento solo se han comparado los datos de la población de la ciudad, vamos a extender estos a la población de toda la provincia. Haremos también la distinción entre la población total y solo la comprendida entre los 14 y 54 años, para ver como esto afecta a los datos.

En primer lugar, y para seguir el mismo orden que se ha utilizado antes, vamos a comparar los datos de población total de la provincia.

➢ TABLA 14:POBLACIÓN PROVINCIA/USUARIOS

Localización	Población	Usuarios	Población/usuarios
Madrid	6454440	-	-
Barcelona	5523784	70	78911.2
Valencia	2548898	40	63722.45
Sevilla	1941355	38	51088.29
Zaragoza	960111	14	68579.36
Málaga	1621968	34	47704.94
Murcia	1466818	24	61117.42
Bilbao	1151905	22	52359.32
Alicante	1868438	30	62271.27
Córdoba	799402	12	66616.83
Valladolid	529157	28	18898.46
Vigo	950919	16	59432.44
Gijón	1061756	18	58986.44
La Coruña	1132735	26	43566.73
Vitoria	321932	14	22995.14
Granada	919455	38	24196.18
Oviedo	1061756	26	40836.77

Como vemos en la tabla anterior, las relaciones entre población y usuarios a cambiado ligeramente. Entre los datos más destacados podemos ver como el valor de Zaragoza ahora entra dentro de los valores comunes de las otras ciudades. Otros datos a tener en cuenta son los datos de Vitoria, Granada y Valladolid que ahora se encuentran muy por debajo de la media. Esto puede deberse a varios factores:

- El primero de ellos puede ser debido a la baja población de las provincias que estamos comparando. Debido a que la precisión de los experimento es baja, como comentamos anteriormente son múltiplos de 10 usuarios, esto puede haber ocasionado que el número de usuarios se viera ampliamente redondeado hacia arriba en estas localizaciones.

- El segundo factor son las horas y días en los que se realizó el experimento. Este hecho, que se comento previamente como uno de los factores clave a la hora de tener en cuenta los valores obtenidos de cada uno de los experimentos, puede haber ocasionado que el valor medio de los usuarios necesarios en estas zonas geográficas se haya visto incrementado.

Como hemos hecho previamente, vamos a comparar también los datos de usuarios con las poblaciones de cada una de las provincias comprendidas entre 14 y 54 años. Viendo que los datos que se han obtenido en la anterior tabla han mejorado bastante las relaciones entre las distintas zonas, cabe esperar que los datos de esta nueva tabla sean aun mejores.

➢ TABLA 15:POBLACION PROVINCIA ACTIVA/USUARIOS

Localización	Población	Usuarios	Población/usuarios
Madrid	4688965	-	-
Barcelona	3902654	70	55752.2
Valencia	1808070	40	45201.75
Sevilla	1441526	38	37934.89
Zaragoza	652587	14	46631.36
Málaga	1179253	34	34683.91
Murcia	1105050	24	46043.75
Bilbao	754624	22	34301.09
Alicante	1296053	30	43201.77
Córdoba	565396	12	47116.33
Valladolid	351534	28	12556.78
Vigo	638140	16	39883.75
Gijón	658560	18	36586.67
La Coruña	727127	26	27966.42
Vitoria	219723	14	15694.5
Granada	662860	38	17443.68
Oviedo	658560	26	25329.23

Como vemos en la tabla, aun está presente el problema que comentábamos previamente en las localizaciones de Valladolid Vitoria y Granada, aunque este se ha visto algo reducido, siendo menos que los demás datos, pero no demasiado.

Centrándonos ahora en el conjunto de datos podemos apreciar varias características:

- Los datos de las ciudades que tienen una localización de Trending Topic en Twitter son bastante parejas, solo el caso de Barcelona destaca por ser algo mayor que el de las demás.

- Los datos de las demás ciudades, exceptuando las previamente comentadas, son muy parecidos, variando su relación levemente.

De los datos previamente expuestos vemos como el número de cuentas necesarias depende de la población aproximadamente con un factor de 1 usuario=40000 habitantes de tal forma que en Madrid serían necesarios 118 usuarios para conseguir que apareciera el hashtag como tema del momento. Esto tiene sentido ya que, como hemos visto en la tabla de inicio de este apartado, no se consiguió en ningún momento que apareciera el hashtag utilizado.

6.2. Generación de un Trending Topic.

Una vez hemos visto el experimento previo donde probábamos en cada una de las geolocalizaciones para descubrir cuál sería el número mínimo de usuarios necesarios para aparecer en la página de trendsmap, vamos a utilizar dicha información para intentar conseguir ser Trending Topic en Twitter.

Para ello, dentro de las geolocalizaciones que probamos en el caso anterior, cogeremos los datos de todas aquellas que tienen un número asociado de woeid en Twitter, para que esta pueda resultar como Trending Topic. De la tabla que tenemos del apartado anterior nos quedaría:

➢ TABLA 16:RESULTADO DEL EXPERIMENTO EN CIUDADES CON TT

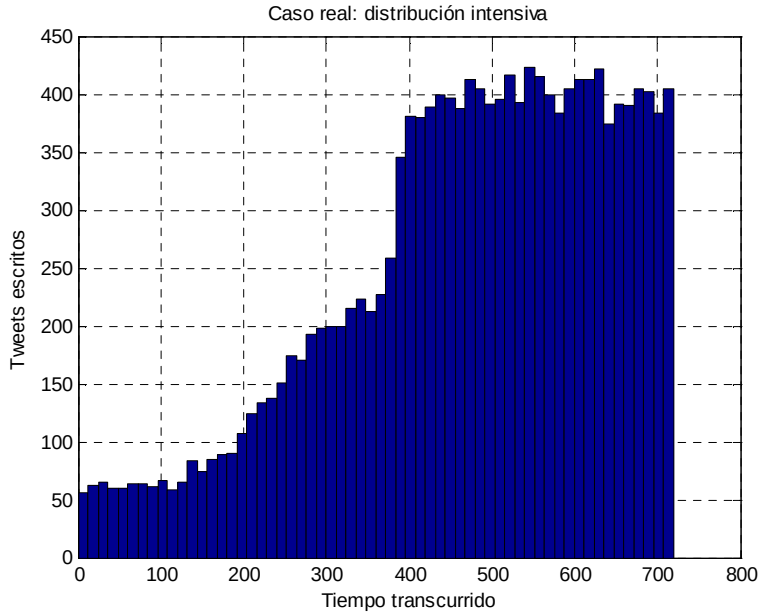
Experimento / Localización	#1	#2	#3	#4	#5	Media	Desviación típica
Murcia	20	30	30	20	20	24	5.48
Málaga	30	30	20	50	40	34	11.40
Sevilla	40	40	40	30	40	38	4.47
Valencia	50	40	40	30	40	40	7.07
Zaragoza	20	20	10	10	10	14	5.48
Bilbao	30	20	20	20	20	22	4.47
Barcelona	70	60	80	70	70	70	7.07
Madrid	-	-	-	-	-	-	-

Como vemos, las zonas donde el número de usuarios es menor es en Zaragoza, Murcia y Bilbao, por lo que probaremos en las tres localizaciones para intentar conseguir nuestro objetivo. De esta forma estamos maximizando la probabilidad de que nuestro experimento consiga el objetivo deseado.

Los experimentos para intentar ser Trending Topic utilizarán la distribución intensiva, vista en el apartado 4.1.1, ya que esta es la mejor distribución para realizar pruebas de larga duración, ya que mantiene durante más tiempo el máximo de publicación.

La duración de estos varía entre dos valores, 12 y 24 horas. En ambos casos lo que se intenta es publicar durante la noche, ya que es cuando menos gente esta activa dentro de Twitter.

Una vez hemos visto las características básicas de este tipo de experimentos, vamos a ver una de las numerosas ejecuciones realizadas.



➢ ILUSTRACIÓN 33:EJEMPLO PRÁCTICO DE PUBLICACIÓN INTENSIVA

Este experimento en concreto se trata de una ejecución de 12 horas que se realizó en Zaragoza. Como vemos, la distribución generada se parece mucho a la simulación con retweets que se generó para probar el correcto funcionamiento de la distribución.

A pesar de los numerosos intentos que se han realizado en las distintas zonas comentadas previamente, ninguno de los experimento logró aparecer como Trending Topic en Twitter.

Como hemos visto durante este apartado generar un Trending Topic es una tarea muy difícil, ya que pese al haber utilizado un gran número de usuarios no se ha conseguido el objetivo. Esto puede ser debido a que Twitter no solo tenga en cuenta la geolocalización asociada a cada uno de los Tweets generados sino que Twitter también tenga en cuenta la IP desde la cual se envío el mensaje. Siendo de esta forma, tendría más sentido que nuestro hashtag apareciera en la página web y no en la Twitter, ya que la página Trendsmap no tendría acceso a la IP de los usuarios desde la cual se ha enviado el mensaje.

7. Conclusiones.

Como hemos visto en la introducción del proyecto, teníamos dos objetivos principales, el estudio de usuarios necesarios en función de cada zona geográfica y conseguir ser Trending Topic en alguna de estas zonas.

Para cumplir estos objetivos se ha diseñado una herramienta capaz de, a partir de una palabra que será el hashtag que se utilizará durante el experimento, generar mensajes aleatorios y publicarlos en las distintas geolocalizaciones. Además, para ver si la herramienta a cumplido su objetivo esta analiza tanto en Twitter como en la página trendsmap.com si el mensaje que se utiliza como hashtag ha logrado aparecer como tendencia del momento.

Hemos realizado dos tipos de experimentos para realizar nuestros objetivos. En primer lugar se han realizado experimentos en cada una de las geolocalizaciones para determinar el número de usuarios mínimo en cada una de las zonas. Posteriormente, una vez analizados estos datos se realizó el segundo experimento donde se utilizaban todos los usuarios virtuales disponibles para intentar ser Trending Topic. Estos dos experimentos tenían características diferentes, distribución y tiempo de vida, de tal forma que se adaptarán mejor al objetivo que intentaban cumplir.

En los experimentos hemos visto que el volumen de datos necesario para las distintas geolocalizaciones no varía mucho, si lo comparamos con la población activa en las redes sociales y con la población provincial y no solo la municipal. Como vemos estas zonas designadas por Twitter no solo cubren el área metropolitano sino que cubren un gran área entorno a estos. Además, hemos conseguido analizar y sacar una relación entre el número de usuarios activos y las cuentas necesarias para que nuestra herramienta apareciera entre las tendencias más habladas. Este factor es de $1\ usuario = 40000\ habitantes$ que se podría utilizar como punto de partida para futuros experimentos en otras zonas geográficas.

Como se ha visto a lo largo del proyecto, se han debido tener en cuenta las reglas que impone Twitter a sus usuarios en cuanto a spam, ya que al automatizar el uso de los usuarios virtuales debíamos incluir un sistema capaz de generar frases aleatorias para cada uno de ellos. Además, dentro del segundo objetivo, conseguir ser Trending Topic, no solo necesitábamos las cuentas virtuales sino que también se necesitarán distintas IPs para las cuentas.

8.Líneas futuras.

En cuanto a las mejoras propuestas para el sistema previamente descrito existen varias:

- La primera es la mejora de la inteligencia artificial utilizada para generar las frases que posteriormente se publicarán. Actualmente, como se ha descrito previamente, la generación de lenguaje se realiza de una forma totalmente aleatoria, cogiendo palabras al azar de la base de datos para cumplir el patrón propuesto de frase. Este sistema se podría mejorar implementando una IA capaz de elegir correctamente las palabras que queremos utilizar en la frase de tal forma que esto tuviera sentido. De esta forma las frases generadas para el experimento se parecerían más a las que puede generar un ser humano.

- La segunda mejora se corresponde a la interacción entre cuentas del experimento. Actualmente la única interacción que desempeñan las cuentas del experimento es la labor de realizar retweets a mensajes previamente enviados por otras cuentas. Esto se podría mejorar haciendo que las cuentas se nombraran entre sí, de tal forma que se crearan conversaciones virtuales entre ellas. Este sistema se podría hacer reactivo, es decir, cada vez que una cuenta enviará un mensaje a otra, esta avisaría a la otra de que le ha enviado un mensaje, de tal forma que el tiempo de espera para la siguiente publicación de la cuenta nombrada sería menor del original. Estas conversaciones entre cuentas se podrían definir mediante grupos de usuarios que se nombraran entre ellos de forma aleatoria, y estos grupos se podrían definir estática (al inicio de la ejecución del programa) o dinámica, donde los grupos irían variando el número de usuarios activos. Esta mejora unida a una mejora de la IA del contenido de los usuarios pueden hacer que las conversaciones no solo se llevarán a cabo entre usuarios del experimento sino que incluso se podría utilizar como forma de interacción con cuentas reales.

- Otra mejora para implementar sería la inserción de contenido multimedia con los mensajes generados por los usuarios, de tal forma que el contenido enviado a Twitter tuviera una mayor relevancia en la red social.

9.Bibliografía

A continuación se muestran los enlaces utilizados como referencia tanto como para la documentación previa al proyecto como para el propio desarrollo de este:

[1] **Redes Sociales - Definición de redes sociales.**
http://recursostic.educacion.es/observatorio/web/es/internet/web-20/1043-redes-sociales?start=1

[2] **Entrada de wikipedia en inglés de redes sociales:**
https://en.wikipedia.org/wiki/Social_network

[3] **Historia de las redes sociales:**
http://recursostic.educacion.es/observatorio/web/es/internet/web-20/1043-redes-sociales?start=2

[4] **Entrada de wikipedia en inglés de Geocities:**
https://en.wikipedia.org/wiki/GeoCities

[5] **Entrada de wikipedia en inglés de Friendster:**
https://en.wikipedia.org/wiki/Friendster

[6] **Entrada de wikipedia en inglés de Facebook:**
https://en.wikipedia.org/wiki/Facebook

[7] **Entrada de wikipedia en inglés de Twitter:**
http://en.wikipedia.org/wiki/Twitter

[8] **Usuarios activos en Twitter:**
http://www.adweek.com/socialtimes/twitter-300-million-mau/500394

[9] **Twitter4j:**
http://twitter4j.org/en/index.html

[10] **Entrada de wikipedia en inglés de SPAM:**
http://en.wikipedia.org/wiki/Spamming

[11] **Economía del Spam:**
http://blogs.wsj.com/ideas-market/2012/08/13/the-economics-of-spam/

[12] **Promoted Trending Topics**
https://business.twitter.com/es/help/what-are-promoted-trends?location=emea

[13] **Técnicas de spam**
http://www.securelist.com/en/threats/spam?chapter=95

[14] **Técnicas de spam**
http://www.prismemail.com/abouttechniques.php

[15] **Programas antispam:**
http://www.anti-spam.com.es/programas

[16] **Spam en las redes sociales**
http://mashable.com/2013/09/30/social-media-spam-study/

[17] **Spam en Twitter**
http://jacarballar.wordpress.com/2012/10/30/que-es-el-spam-de-twitter-y-su-relacion-con-el-secuestro-de-cuentas/
[18] **Análisis del Spam en Twitter:**
http://conferences.sigcomm.org/imc/2011/docs/p243.pdf
[19] **Políticas de seguridad en Twitter:**
https://support.twitter.com/groups/56-policies-violations/topics/236-twitter-rules-policies/articles/72688-las-reglas-de-twitter
[20] **Trackgirl**
http://www.wired.com/2012/06/twitter_arm/
[21] **Información sobre NLG:**
http://web.science.mq.edu.au/~rdale/publications/papers/1997/jnle97.pdf
[22] **SimpleNLG:**
https://code.google.com/p/simplenlg/
[23] **RFC OAuth:**
https://tools.ietf.org/html/rfc6749
[24] **OAuth en Twitter:**
https://dev.twitter.com/oauth/overview
[25] **Entrada de wikipedia en inglés de geolocalización:**
http://en.wikipedia.org/wiki/Geolocation
[26] **Página oficial de trendsmap:**
http://trendsmap.com/
[27] **Instituto Nacional de Estadísticas:**
http://www.ine.es/
[28] **Twitter developers:**
https://dev.twitter.com/
[29] **Edad media de uso de redes sociales:**
http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/
[30] **Teoría de los Seis grados de separación**
https://es.wikipedia.org/wiki/Seis_grados_de_separaci%C3%B3n
[31] **¿Qué son las etiquetas?**
https://support.twitter.com/articles/247830-que-son-las-etiquetas-simbolos
[32] **Preguntas frecuentes sobre las tendencias en Twitter**
https://support.twitter.com/articles/349215-preguntas-frecuentes-sobre-las-tendencias-en-twitter
[33] **Social media spam**
http://nexgate.com/wp-content/uploads/2013/09/Nexgate-2013-State-of-Social-Media-Spam-Research-Report.pdf
[34] **Spam decreased in Twitter**
https://blog.twitter.com/2010/state-twitter-spam

[35] **Trending Topic spam**
http://www.bernardosignes.com/cuando-el-trending-topic-es-spam
[36] **Salida a Bolsa de Twitter**
http://www.eleconomista.es/mercados-cotizaciones/noticias/5293481/11/13/Twitter-marca-la-segunda-mejor-salida-a-bolsa-en-la-historia-de-Internet.html#.Kku8mrTuhpcAfsH
[37] **Salida a Bolsa de Facebook**
http://www.rtve.es/noticias/20120518/facebook-espera-captar-15000-millones-euros-su-salida-hoy-bolsa/528277.shtml
[38] **Evolución de los usuarios en las redes sociales**
http://www.clasesdeperiodismo.com/2015/01/10/pew-research-facebook-sigue-siendo-el-rey/
[39] **Sophos Security Threat Report reveals increase in social networking security threats**
https://www.sophos.com/en-us/press-office/press-releases/2011/01/threat-report-2011.aspx

10. Anexos

10.1. Presupuesto

En este anexo mostraremos los costes asociados a la realización del proyecto para desempeñar la herramienta de publicación en Twitter. Como veremos, el coste de la herramienta no es demasiado alto ya que la mayor parte del presupuesto viene determinado por el personal involucrado en el proyecto, ya que las licencias utilizadas en este son de carácter gratuito.

Autor

Alberto Chicharro Sobrino

Departamento

Departamento de Telemática

Descripción del proyecto

TÍTULO: Generación de Trending Topic artificiales.

DURACIÓN: 10 meses.

TASA DE COSTES INDIRECTOS: 20%.

Personal

En la siguiente tabla se muestra el coste asociado al personal involucrado dentro del proyecto:

Apellidos y nombre	Categoría	Dedicación(meses)	Coste por mes	Coste Total
Chicharro Sobrino, Alberto	Ingeniero	9	2694.39€	24249.51€
Carrascosa Amigo, Juan Miguel	Ingeniero Senior	1	4.289,54€	4289.54€
Cuevas Rumin, Rubén	Ingeniero Senior	1	4.289,54€	4289.54€

Equipos

Además del coste de personal, tenemos que tener en cuenta el coste asociado al equipo necesario para el desarrollo del proyecto:

Descripción	Coste	% Uso en el proyecto	Dedicación (meses)	Periodo de deprecación	Coste imputable
Ordenador de pruebas	500,00€	100	10	60	83.33€
Servidor de pruebas	800,00€	100	10	60	133.33€

Subcontratación de tareas

Ya que el proyecto se ha realizado completamente por el personal previamente documentado, los costes de subcontratación son nulos.

Otros costes directos del proyecto

Costes directos asociados a el desarrollo del proyecto:

Descripción	Empresa	Coste
Fungible	-	200,00€
Viajes	-	100,00€

Coste total del proyecto

Por tanto, una vez hemos visto el desglose de costes del proyecto, vamos a ver el valor total del desarrollo del proyecto:

Descripción	Coste
Personal	32.828,59€
Equipos	216,66€
Subcontratación	0,00€
Otros	300,00€
Costes indirectos	5.811,14€
Total	39.156,39€